

Improving Sparse Representation-Based Classification Using Local Principal Component Analysis

Chelsea Weaver*, Naoki Saito

*Department of Mathematics
University of California, Davis
One Shields Avenue
Davis, California, 95616, United States*

Abstract

Sparse representation-based classification (SRC), proposed by Wright et al., seeks the sparsest decomposition of a test sample over the dictionary of training samples, with classification to the most-contributing class. Because it assumes test samples can be written as linear combinations of their same-class training samples, the success of SRC depends on the size and representativeness of the training set. Our proposed classification algorithm enlarges the training set by using local principal component analysis to approximate the basis vectors of the tangent hyperplane of the class manifold at each training sample. The dictionary in SRC is replaced by a local dictionary that adapts to the test sample and includes training samples and their corresponding tangent basis vectors. We use a synthetic data set and three face databases to demonstrate that this method can achieve higher classification accuracy than SRC in cases of sparse sampling, nonlinear class manifolds, and stringent dimension reduction.

Keywords: sparse representation, local principal component analysis, dictionary learning, classification, face recognition, class manifold

2016 MSC: 00-01, 99-00

1. Introduction

We are concerned with *classification*, the task of assigning labels to unknown samples given the class information of a training set. Some practical applications of classification include the recognition of handwritten digits [1] and face recognition [2, 3, 4]. These tasks are often very challenging. For example, in face recognition, the classification algorithm must be robust to within-class variation in properties such as expression, face/head angle, changes in hair or makeup, and differences that may occur in the image environment, most notably, the lighting conditions [4]. Further, in real-world settings, we must be able to handle greatly-deficient training data (i.e., too few or too similar training samples, in the sense that the

*Corresponding author

Email addresses: `caweaver@math.ucdavis.edu` (Chelsea Weaver), `saito@math.ucdavis.edu` (Naoki Saito)

given training set is insufficient to generalize the data set’s class structure) [5], as well as occlusion and noise [2].

In 2009, Wright et al. proposed *sparse representation-based classification* (SRC) [2]. SRC was motivated by the recent boom in the use of sparse representation in signal processing (see, e.g., the work of Candes [6]). The catalyst of these advancements was the discovery that, under certain conditions, the sparsest representation of a signal using an over-complete set of vectors (often called a *dictionary*) could be found by minimizing the ℓ^1 -norm of the representation coefficient vector [7]. Since the ℓ^1 -minimization problem is convex, this gave rise to a tractable approach to obtaining the sparsest solution.

SRC applies this relationship between the minimum ℓ^1 -norm and the sparsest solution to classification. The algorithm seeks the sparsest decomposition of a test sample over the dictionary of training samples via ℓ^1 -minimization, with classification to the class whose corresponding portion of the representation approximates the test sample with least error. The method assumes that class manifolds are linear subspaces, so that the test sample can be represented using training samples in its ground truth class. Wright et al. [2] argue that this is precisely the sparsest decomposition of the test sample over the training set. They make the case that sparsity is critical to high-dimensional image classification and that, if properly harnessed, it can lead to superior classification performance, even on highly corrupted or occluded images. Further, good results can be achieved regardless of the choice of image features that are used for classification, provided that the number of retained features is large enough [2]. Though SRC was originally applied to face recognition, similar methods have been employed in clustering [8], dimension reduction [9], and texture and handwritten digit classification [10].

The SRC assumption that class manifolds are linear subspaces is often violated; e.g., facial images that vary in pose and expression are known to lie on nonlinear class manifolds [11, 12]. Additionally, small training set size, one of the primary challenges in face recognition and classification as a whole, can easily make it impossible to represent a given test sample using its same-class training samples, even in the case that the class manifold is linear. However, these reasons alone are not enough to discount SRC even on such data sets, as demonstrated by Wright et al. [2] in experiments on the AR face database [13]. AR contains expression and occlusion variations that suggest the underlying class manifolds are nonlinear, yet SRC often outperformed SVM (support vector machines) on AR for a wide variety of feature extraction methods and feature dimensions [2]. To understand how this is possible, consider that SRC decomposes the test sample over the entire training set, and so components of the test sample not within the span of its ground truth class’s training samples may be absorbed by training samples from other classes. A similar fail-safe occurs when the class manifolds (linear or otherwise) are sparsely sampled.

The above discussion, however, illustrates a weakness in SRC. When the algorithm relies on “wrong-class” training samples to partially represent or approximate the test sample, misclassification may ensue,

especially when the class manifolds are close together. In the case where class manifolds are nonlinear and/or sparsely sampled, so that it is impossible to accurately approximate the test sample using only the training samples in its ground truth class, this approximation could conceivably be improved if we were able to increase the sampling density around the test sample, “fleshing out” its local neighborhood on the (correct) class manifold. This is the motivation behind this paper’s proposed classification algorithm.

Our contributions in this paper are the following:

1. We introduce a classification algorithm that improves SRC by increasing the accuracy and locality of the approximation of the test sample in terms of its ground truth class. Our algorithm is designed to increase the training set via nearby (to the test sample) basis vectors of the hyperplanes approximately tangent to the (unknown) class manifolds. This provides the two-fold benefit of counter-balancing the potential sparse sampling of class manifolds (especially in the case that they are nonlinear) and helping to retain more information in few dimensions when used in conjunction with dimension reduction.
2. We state guidelines for the setting of parameters in this algorithm and analyze its computational complexity and storage requirements.
3. We demonstrate that our algorithm leads to classification accuracy exceeding that of traditional SRC and related methods on a synthetic database and three popular face databases. We thoroughly analyze and explain our experimental results (e.g., accuracy, runtime, and dictionary size) of the compared algorithms.
4. We illustrate that the tangent hyperplane basis vectors used in our method can capture sample details lost during principal component analysis in the case of face recognition.

This paper is organized as follows: In Section 2, we discuss work related to our proposed method, and we state SRC in detail in Section 3. In Section 4, we describe our proposed classification algorithm and discuss its parameters, computational complexity, and storage requirements. We present our experimental results in Section 5, and in Section 6, we summarize our findings and discuss avenues of future work.

Setup and Notation. We assume that the input data is represented by vectors in \mathbb{R}^m and that dimension reduction, if used, has already been applied. The training set, i.e., the matrix whose columns are the data samples with known class labels, is denoted by $X_{\text{tr}} = [\mathbf{x}_1, \dots, \mathbf{x}_{N_{\text{tr}}}] \in \mathbb{R}^{m \times N_{\text{tr}}}$. The number of classes is denoted by $L \in \mathbb{N}$, and we assume that there are N_l training samples in class l , $1 \leq l \leq L$. Lastly, we refer to a given test sample by $\mathbf{y} \in \mathbb{R}^m$.

2. Related Work

The approach of using tangent hyperplanes for pattern recognition is not new. When the data is assumed to lie on a low-dimensional manifold, local tangent hyperplanes are a simple and intuitive approach to enhancing the data set and gaining insight into the manifold structure. Our proposed method is very much

related to *tangent distance classification* (TDC) [14, 15, 16], which constructs local tangent hyperplanes of the class manifolds, computes the distances between these hyperplanes and the given test sample, and then classifies the test sample to the class with the closest hyperplane. We show in Section 5 that our proposed method’s integration of tangent hyperplane basis vectors into the sparse representation framework generally outperforms TDC.

On the other hand, approaches to address the limiting linear subspace assumption (i.e., the assumption that class manifolds are linear subspaces) in SRC have been proposed. For example, Ho et al. extended sparse coding and dictionary learning to general Riemannian manifolds [17]. Admittedly only a first step in meeting their ultimate objective, Ho et al.’s work requires explicit knowledge of the class manifolds. This is an unsatisfiable condition in many real-world classification problems and is not a requirement of our proposed algorithm. Alternatively, *kernel methods* have been effective in overcoming SRC’s linearity assumption, as nonlinear relationships in the original space may be linear in kernel space given an appropriate choice of kernel [18]. The method proposed in *kernel collaborative face recognition*, for example, by Wang et al. [19], uses the kernel trick with the Hamming kernel and the local binary patterns of facial images. This method was shown to offer a substantial performance improvement (in terms of both accuracy and runtime) over SRC. Our proposed method is likely kernelizable; this and its use in conjunction with customized or critically-extracted features (such as local binary patterns) are not investigated in this paper.

Several “local” modifications of SRC implicitly ameliorate the linearity assumption; in *collaborative neighbor representation-based classification* [20] and *locality-sensitive dictionary learning* (LSDL-SRC) [21], for instance, coefficients of the representation are constrained by their corresponding training samples’ distances to the test sample, and so these algorithms need only assume linearity at the local level. Our proposed method is designed to improve not only the locality but also the accuracy of the approximation of the test sample in terms of its ground truth class. Section 5 contains an experimental comparison between our proposed method and LSDL-SRC, as well as a discussion thereof.

Other classification algorithms have been proposed that are similar to ours in that they aim to enlarge or otherwise enhance the training set in SRC. Such methods for face recognition, for example, include the use of virtual images that exploit the symmetry of the human face, as in both the method of Xu et al. [22] and *sample pair based sparse representation classification* [23]. Though visual comparison of these virtual images and our recovered tangent vectors (see Section 5.4.6) could be informative, our proposed method can be used for general classification. As an alternative approach, a *multi-task joint sparse representation* (MJSR) framework has been used to improve the classification accuracy of SRC. Yuan et al. [24] defined these multiple tasks using different modalities of features for successful visual classification, and Yuan et al. [25] combined MJSR with band selection and stepwise Markov random field optimization for hyperspectral image classification. We note that our tangent vector approach is fundamentally different from that of multi-task

learning. Further, our proposed algorithm could be amended to the MJSR framework.

Additionally, there have been many local modifications to the sparse representation framework with objectives other than classification. For example, Li et al.’s *robust structured subspace learning* (RSSL) [26] uses the $\ell_{2,1}$ -norm for sparse feature extraction, combining high-level semantics with low-level, locality-preserving features. In the feature selection algorithm *clustering-guided sparse structural learning* (CGSSL) by Li et al. [27], features are jointly selected using sparse regularization (via the $\ell_{2,1}$ -norm) and a non-negative spectral clustering objective. Not only are the selected features sparse; they also are the most discriminative features in terms of predicting the cluster indicators in both the original space and a lower-dimensional subspace on which the data is assumed to lie.

3. Sparse Representation-Based Classification

SRC [2] solves the optimization problem

$$\boldsymbol{\alpha}^* := \arg \min_{\boldsymbol{\alpha} \in \mathbb{R}^{N_{\text{tr}}}} \|\boldsymbol{\alpha}\|_1, \text{ subject to } \mathbf{y} = X_{\text{tr}}\boldsymbol{\alpha}. \quad (1)$$

It is assumed that the training samples have been normalized to have ℓ^2 -norm equal to 1, so that the representation in Eq. (1) will not be affected by the samples’ magnitudes. The use of the ℓ^1 -norm in the objective function is designed to approximate the ℓ^0 -“norm,” i.e., to aim at finding the smallest number of training samples that can accurately represent the test sample \mathbf{y} . It is argued that the nonzero coefficients in the representation will occur primarily at training samples in the same class, so that

$$\text{class_label}(\mathbf{y}) = \arg \min_{1 \leq l \leq L} \|\mathbf{y} - X_{\text{tr}}\delta_l(\boldsymbol{\alpha}^*)\|_2 \quad (2)$$

produces the correct class assignment. Here, δ_l is the indicator function that acts as the identity on all coordinates corresponding to samples in class l and sets the remaining coordinates to zero. In other words, \mathbf{y} is assigned to the class whose training samples contribute the most to the sparsest representation of \mathbf{y} over the entire training set.

The reasoning behind this is the following: It is assumed that the class manifolds are linear subspaces, so that if each class’s training set contains a spanning set of the corresponding subspace, the test sample can be expressed as a linear combination of training samples in its ground truth class. If the number of training samples in each class is small relative to the number of total training samples N_{tr} , this representation is naturally sparse [2].

As real-world data is often corrupted by noise, the constrained ℓ^1 -minimization problem in Eq. (1) may be replaced with its regularized version

$$\boldsymbol{\alpha}^* := \arg \min_{\boldsymbol{\alpha} \in \mathbb{R}^{N_{\text{tr}}}} \left\{ \frac{1}{2} \|\mathbf{y} - X_{\text{tr}}\boldsymbol{\alpha}\|_2^2 + \lambda \|\boldsymbol{\alpha}\|_1 \right\}. \quad (3)$$

Here, λ is the trade-off between error in the approximation and the sparsity of the coefficient vector. We summarize SRC in Algorithm 1.

Algorithm 1 Sparse Representation-Based Classification (SRC) [2]

Input: Matrix of training samples $X_{\text{tr}} \in \mathbb{R}^{m \times N_{\text{tr}}}$; test sample $\mathbf{y} \in \mathbb{R}^m$; number of classes L ; and error/sparsity trade-off λ (optional)

Output: The computed class label of \mathbf{y} : $\text{class_label}(\mathbf{y})$

- 1: Normalize each column of X_{tr} to have ℓ^2 -norm equal to 1.
 - 2: Use an ℓ^1 -minimization algorithm to solve either the constrained problem (1) or the regularized problem (3).
 - 3: **for** each class $l = 1, \dots, L$, **do**
 - 4: Compute the norm of the class l residual: $\text{err}_l(\mathbf{y}) := \|\mathbf{y} - X_{\text{tr}}\delta_l(\boldsymbol{\alpha}^*)\|_2$.
 - 5: **end for**
-

4. Proposed Algorithm

4.1. Local Principal Component Analysis Sparse Representation-Based Classification

Our proposed algorithm, *local principal component analysis sparse representation-based classification* (LPCA-SRC), is essentially SRC with a modified dictionary. This dictionary is constructed in two steps: In the offline phase of the algorithm, we generate new training samples as a means of increasing the sampling density. Instead of the linear subspace assumption in SRC, we assume that class manifolds are well-approximated by local tangent hyperplanes. To generate new training samples, we approximate these tangent hyperplanes at individual training samples using *local principal component analysis* (local PCA), and then add the basis vectors of these tangent hyperplanes (after randomly-scaling and shifting them as described in Step 12 of Algorithm 2 and explained in Section 4.3.3) to the original training set. Naturally, the shifted and scaled tangent hyperplane basis vectors (hereon referred to as “tangent vectors”) inherit the labels of their corresponding training samples. The result is an amended dictionary over which a generic test sample can ideally be decomposed using samples that approximate a local patch on the correct class manifold. In the case that the class manifolds are sparsely sampled and/or nonlinear, this allows for a more accurate approximation of \mathbf{y} using training samples (and their computed tangent vectors) from the test sample’s ground truth class. Even in the case that class manifolds are linear subspaces, this technique ideally increases the sampling density around \mathbf{y} on its (unknown) class manifold so that it may be expressed in terms of *nearby* samples.

In the online phase of LPCA-SRC, this extended training set is “pruned” relative to the given test sample, increasing computational efficiency and the locality of the resulting dictionary. Training samples (along with their tangent vectors) are eliminated from the dictionary if their Euclidean distances to the given test sample are greater than a threshold, and then classification proceeds as in SRC as the test sample is sparsely decomposed (via ℓ^1 -minimization) over this local dictionary.

The method in LPCA-SRC has an additional benefit: When SRC is applied to the classification of high-resolution images (e.g., $> O(10^4)$ pixels), some method of dimension reduction is generally necessary to reduce the dimension of the raw samples, due to the high computational complexity of solving the ℓ^1 -minimization problem. Basic dimension reduction methods, such as *principal component analysis* (PCA), may result in the loss of class-discriminating details when the PCA feature dimension is small. In Section 5.4.6, we show that the tangent vectors computed in LPCA-SRC can contain details of the raw images that have been lost in the dimension reduction process.

We formally state the offline and online portions of our proposed algorithm in Algorithms 2 and 3, respectively. Obviously, by the definition of “offline phase,” the tangent vectors need only be computed once for any number of test samples. More details regarding the user-set parameters d , n and λ are provided in Sections 4.3.1 and 4.3.2, and an explanation of the pruning parameter r and the tangent vector scaling factor c (in Step 12 of Algorithm 2) are given in Section 4.3.3.

4.2. Local Principal Component Analysis

In LPCA-SRC (in particular, Step 6 of Algorithm 2), we use the local PCA technique of Singer and Wu [28] to compute the tangent hyperplane basis $U^{(l,i)}$. We outline our implementation of their method in Algorithm 4. It computes a basis for the tangent hyperplane $T_{\mathbf{x}_i}\mathcal{M}$ at a point \mathbf{x}_i on the manifold \mathcal{M} , where it is assumed that the local neighborhood of \mathbf{x}_i on \mathcal{M} can be well-approximated by a tangent hyperplane of some dimension $d < m$. A particular strength of Singer and Wu’s method is the weighting of neighbors by their Euclidean distances to the point \mathbf{x}_i , so that closer neighbors play a more important role in the construction of the local tangent hyperplane.

4.3. Remarks on Parameters

In this subsection, we detail the roles of the parameters in LPCA-SRC and suggest strategies for estimating those that must be determined by the user.

4.3.1. Estimate of class manifold dimension and number of neighbors

Recall that d is the estimated dimension of each class manifold and n is the number of neighbors used in local PCA. Both parameters must be inputted by the user in our proposed algorithm. The number of samples in the smallest training class, denoted $N_{l_{\min}}$, limits the range of values for d and n that may be used. Specifically,

$$1 \leq d \leq n < N_{l_{\min}} - 1. \quad (6)$$

This follows from the fact that each training sample must have at least $n + 1$ neighbors in its own class, with the dimension d of the tangent hyperplane being bounded above by the number of columns n in the weighted matrix of neighbors B_i . It is important to observe that when the classes are small (as is often the

Algorithm 2 Local PCA Sparse Representation-Based Classification (LPCA-SRC): **OFFLINE PHASE**

Input: $X_{\text{tr}} = [\mathbf{x}_1 \dots, \mathbf{x}_{N_{\text{tr}}}] \in \mathbb{R}^{m \times N_{\text{tr}}}$; number of classes L ; local PCA parameters d (estimate of class manifold dimension) and n (number of neighbors)

Output: The normalized extended dictionary $D \in \mathbb{R}^{m \times (N_{\text{tr}}(d+1))}$; pruning parameter r

- 1: Normalize the columns of X_{tr} to have ℓ^2 -norm equal to 1.
 - 2: **for** each class $l = 1, \dots, L$ **do**
 - 3: Let $\mathcal{X}^{(l)}$ be the set of class l training samples contained in X_{tr} .
 - 4: **for** each class l training sample $\mathbf{x}_i^{(l)}$, $i = 1, \dots, N_l$ **do**
 - 5: Use local PCA in Algorithm 4 with set of samples $\mathcal{X}^{(l)}$, selected sample $\mathbf{x}_i^{(l)}$, and parameters d and n to compute a basis $U^{(l,i)} := [\mathbf{u}_1^{(l,i)}, \dots, \mathbf{u}_d^{(l,i)}]$ of the tangent hyperplane approximation to the l th class manifold at $\mathbf{x}_i^{(l)}$. Store $U^{(l,i)}$ and $r_i^{(l)} := \|\mathbf{x}_{i_{n+1}}^{(l)} - \mathbf{x}_i^{(l)}\|_2$, the distance between $\mathbf{x}_i^{(l)}$ and its $(n+1)$ st nearest neighbor in $\mathcal{X}^{(l)} \setminus \mathbf{x}_i^{(l)}$.
 - 6: **end for**
 - 7: **end for**
 - 8: Define the pruning parameter $r := \text{median}\{r_i^{(l)} \mid 1 \leq i \leq N_l, 1 \leq l \leq L\}$.
 - 9: Initialize the extended dictionary $D = \emptyset$.
 - 10: **for** each class $l = 1, \dots, L$ **do**
 - 11: **for** each class l training sample $\mathbf{x}_i^{(l)}$, $i = 1, \dots, N_l$ **do**
 - 12: Set $c := r\gamma$, $\gamma \sim \text{unif}(0, 1)$, and form $\tilde{X}^{(l,i)} := [c\mathbf{u}_1^{(l,i)} + \mathbf{x}_i^{(l)}, \dots, c\mathbf{u}_d^{(l,i)} + \mathbf{x}_i^{(l)}, \mathbf{x}_i^{(l)}] \in \mathbb{R}^{m \times (d+1)}$.
 - 13: Normalize the columns of $\tilde{X}^{(l,i)}$ to have ℓ^2 -norm equal to 1 and add it to the extended dictionary:
 $D = [D, \tilde{X}^{(l,i)}]$.
 - 14: **end for**
 - 15: **end for**
-

Algorithm 3 Local PCA Sparse Representation-Based Classification (LPCA-SRC): **ONLINE PHASE**

Input: Test sample $\mathbf{y} \in \mathbb{R}^m$; normalized extended dictionary D ; pruning parameter r ; estimate of class manifold dimension d ; error/sparsity trade-off λ (optional)

Output: The computed class label of \mathbf{y} : $\text{class_label}(\mathbf{y})$.

- 1: Normalize \mathbf{y} to have $\|\mathbf{y}\|_2 = 1$.
- 2: Initialize the pruned dictionary $D_{\mathbf{y}} = \emptyset$ and set $N_{\mathbf{y}} = 0$ (# of columns of $D_{\mathbf{y}}$).
- 3: **for** each class $l = 1, \dots, L$ **do**
- 4: **for** each class l training sample $\mathbf{x}_i^{(l)}$, $i = 1, \dots, N_l$ **do**
- 5: **if** $\|\mathbf{y} - \mathbf{x}_i^{(l)}\|_2 \leq r$ or $\|\mathbf{y} - (-\mathbf{x}_i^{(l)})\|_2 \leq r$ **then**
- 6: Add the portion $\tilde{X}^{(l,i)}$ of D corresponding to $\mathbf{x}_i^{(l)}$ and its tangent vectors to the pruned dictionary:
 $D_{\mathbf{y}} = [D_{\mathbf{y}}, \tilde{X}^{(l,i)}]$. Assign the columns of $\tilde{X}^{(l,i)}$ class l labels. Update $N_{\mathbf{y}} = N_{\mathbf{y}} + (d + 1)$.
- 7: **end if**
- 8: **end for**
- 9: **end for**
- 10: Use an ℓ^1 -minimization algorithm to compute the solution to the constrained problem

$$\boldsymbol{\alpha}^* := \arg \min_{\boldsymbol{\alpha} \in \mathbb{R}^{N_{\mathbf{y}}}} \{ \|\boldsymbol{\alpha}\|_1 \text{ s.t. } \mathbf{y} = D_{\mathbf{y}} \boldsymbol{\alpha} \} \quad (4)$$

or the regularized problem

$$\boldsymbol{\alpha}^* := \arg \min_{\boldsymbol{\alpha} \in \mathbb{R}^{N_{\mathbf{y}}}} \left\{ \frac{1}{2} \|\mathbf{y} - D_{\mathbf{y}} \boldsymbol{\alpha}\|_2^2 + \lambda \|\boldsymbol{\alpha}\|_1 \right\}. \quad (5)$$

- 11: **for** each class $l = 1, \dots, L$, **do**
 - 12: Compute the norm of the class l residual: $\text{err}_l(\mathbf{y}) := \|\mathbf{y} - D_{\mathbf{y}} \delta_l(\boldsymbol{\alpha}^*)\|_2$.
 - 13: **end for**
-

Algorithm 4 Local Principal Component Analysis (Local PCA, adapted from Singer and Wu [28])

Input: Set of samples \mathcal{X} ; selected sample $\mathbf{x}_i \in \mathcal{X}$; dimension of tangent hyperplane d ; number of neighbors n

Output: The basis $U^{(l,i)}$ of the approximated tangent hyperplane at the point \mathbf{x}_i

- 1: Find the $n + 1$ nearest neighbors (with respect to Euclidean distance) of \mathbf{x}_i in $\mathcal{X} \setminus \mathbf{x}_i$. Store the n nearest neighbors as columns of the matrix $X_i := [\mathbf{x}_{i_1}, \dots, \mathbf{x}_{i_n}]$ and use the $(n + 1)$ st nearest neighbor to define $\epsilon_{\text{PCA}} := \|\mathbf{x}_{i_{n+1}} - \mathbf{x}_i\|_2^2$.
 - 2: Form the matrix \bar{X}_i by centering the columns of X_i around \mathbf{x}_i : $\bar{X}_i := [\mathbf{x}_{i_1} - \mathbf{x}_i, \dots, \mathbf{x}_{i_n} - \mathbf{x}_i]$.
 - 3: Form a diagonal weight matrix D_i based on the distance between each neighbor and \mathbf{x}_i as follows: Let $D_i(j, j) = \sqrt{K\left(\frac{\|\mathbf{x}_{i_j} - \mathbf{x}_i\|_2}{\sqrt{\epsilon_{\text{PCA}}}}\right)}$, $j = 1, \dots, n$, where K is the Epanechnikov kernel given by $K(u) := (1 - u^2)\chi_{[0,1]}$.
 - 4: Form the weighted matrix $B_i := \bar{X}_i D_i$.
 - 5: Find the first d left singular vectors of B_i using singular value decomposition. Denote these vectors by $\mathbf{u}_1^{(i)}, \dots, \mathbf{u}_d^{(i)}$.
-

case in face recognition), there are few options for the values of d and n per Eq. (6). Thus these parameters may be efficiently set using cross-validation. This was the method we used to set d and n in the experiments in Section 5. We discuss a recommended cross-validation procedure in Section 4.3.2.

Remark 1. Interestingly, when cross-validation is used to set d , we find empirically that d is often selected to be smaller than the (expected) true class manifold dimension. Further, in these cases, increasing d from the selected value (i.e., increasing the number of tangent vectors used) does not significantly increase classification accuracy. We expect that the addition of even a small number of tangent vectors (those indicating the directions of maximum variance on their local manifolds, per the local PCA algorithm) is enough to improve the approximation of the test sample in terms of its ground truth class. Additional tangent vectors are often unneeded. Since the value of d largely affects LPCA-SRC’s computational complexity and storage requirements, these observations suggest that when the true manifold dimension is large, it is better to underestimate it than overestimate it. Further, setting $d = 1$ can often produce a good result, hence $d = 1$ could be used by default.

There are other methods for determining d besides cross-validation and fixing $d = 1$. One may use the multiscale SVD algorithm of Little et al. [29] or Ceruti et al.’s DANCo (*Dimensionality from Angle and Norm Concentration* [30]). However, in our experiments in Section 5, we set d using cross-validation. See Section 4.3.2 below.

Remark 2. Certainly, the parameters d and n could vary per class, i.e., d and n could be replaced with d_l and n_l , respectively, for $l = 1, \dots, L$. In face recognition, however, if each subject is photographed under similar conditions, e.g., the same set of lighting configurations, then we expect that the class manifold dimension

is approximately the same for each subject. Further, without some prior knowledge of the class manifold structure, using distinct d and n for each class may unnecessarily complicate the setting of parameters in LPCA-SRC.

4.3.2. Using cross-validation to set multiple parameters

On data sets of which we have little prior knowledge, it may be necessary to use cross-validation to set multiple parameters in LPCA-SRC. Since grid search (searching through all parameter combinations in a brute-force manner) is typically expensive, we suggest that cross-validation be applied to the parameters n , λ , and d , consecutively in that order as needed.¹ During this process, we recommend holding the error/sparsity trade-off λ (if used) equal to a small, positive value (e.g., $\lambda = 0.001$) and setting $d = 1$ until these parameters' respective values are determined. We justify and detail this approach below.

Our reasons for suggesting this consecutive cross-validation procedure is the following: During experiments, we found that the LPCA-SRC algorithm can be quite sensitive to the setting of n , especially when there are many samples in each training class (since there are many possible values for n). This is expected, as the setting of n affects both the accuracy of the tangent vectors and the pruning parameter r . In contrast, LPCA-SRC is empirically fairly robust to the values of λ and d used, and as mentioned in Remark 1, setting $d = 1$ can result in quite good performance in LPCA-SRC, even when the true dimension of the class manifolds is expected to be larger.

The values tried for d and n during cross-validation must satisfy Eq. (6). They must clearly also be integers. Determining an appropriate set of values over which to cross-validate λ is not so clear-cut; however, this is a challenge whenever a regularized optimization problem (such as SRC's Eq. (3)) is used. We show an example of our cross-validation procedure for LPCA-SRC in Algorithm 5. The example illustrates the case that $N_{l_{\min}} = 5$; note that this is true for the ORL database in our experiments in Section 5.4.5. We also give an example set of values from which to select λ .

Let us use the notation introduced in Algorithm 5 and consider how to determine the sets of values N_{vals} and D_{vals} (over which n and d , respectively, will be cross-validated) if there are more samples in each training class, say $N_{l_{\min}} = 25$. One could set $N_{\text{vals}} = \{1, 2, 5, 10, 15, 20, 23\}$, and, given Remark 1, $D_{\text{vals}} = \{1, 2, 5, 10\}$, for example. *We recommend that N_{vals} be chosen to include no more than 5-10 fairly evenly-spaced values that satisfy Eq. (6), with $D_{\text{vals}} \subseteq N_{\text{vals}}$ (possibly omitting the larger values in N_{vals} , e.g., those greater than 10).* This was our approach in the experiments in Section 5. While the looseness of these guidelines may be unsatisfying, we stress that the performance of LPCA-SRC is robust to the exact set of values used. Choosing a handful of arbitrary values that satisfy Eq. (6) to construct N_{vals} and D_{vals} is sufficient.

¹If the constrained optimization problem (Eq. (4)) is used, the error/sparsity trade-off λ is not needed.

Algorithm 5 EXAMPLE Setting of Parameters in LPCA-SRC

Input: Database training set with smallest class size $N_{l_{\min}} = 5$

Output: The parameters $n = n^*$, $\lambda = \lambda^*$, and $d = d^*$ for use in LPCA-SRC (specifically, n and d for Algorithm 2 and λ for Algorithm 3)

- 1: Define $N_{\text{vals}} = \{1, 2, 3 = N_{l_{\min}} - 2\}$, $D_{\text{vals}} = \{1, 2, 3\}$, $\Lambda_{\text{vals}} = \{0.0001, 0.001, 0.01, 0.05, 0.1, 0.3, 0.5\}$.
 - 2: Use the training set to perform cross-validation on n over N_{vals} with $d = 1$ and $\lambda = \lambda_0$, where λ_0 is randomly selected from Λ_{vals} . Obtain $n = n^*$, where $n^* \in N_{\text{vals}}$ produces the maximum classification accuracy in LPCA-SRC.
 - 3: Similarly, perform cross-validation on λ over Λ_{vals} with $n = n^*$ and $d = 1$. Obtain $\lambda = \lambda^*$.
 - 4: Similarly, perform cross-validation on d over $D_{\text{vals}} \cap \{1, \dots, n^*\}$ with $n = n^*$ and $\lambda = \lambda^*$. Obtain $d = d^*$.
-

4.3.3. Pruning parameter

First, we stress that the pruning parameter r is not a user-set parameter. Its value is automatically computed in the offline phase of LPCA-SRC (Algorithm 2). We explain this process here.

Recall that we only include a training sample $\mathbf{x}_i^{(l)}$ and its tangent vectors in the pruned dictionary $D_{\mathbf{y}}$ if $\mathbf{x}_i^{(l)}$ (or its negative) is in the closed Euclidean ball $\overline{B_m(\mathbf{y}, r)} \subset \mathbb{R}^m$ with center \mathbf{y} and radius r . Thus r is a parameter that prunes the extended dictionary D to obtain $D_{\mathbf{y}}$. A smaller dictionary is good in terms of computational complexity, as the ℓ^1 -minimization algorithm will run faster. Further, we can obtain this computational speedup without (theoretically) degrading classification accuracy: If $\pm \mathbf{x}_i^{(l)}$ is far from \mathbf{y} in terms of Euclidean distance, then it is assumed that $\pm \mathbf{x}_i^{(l)}$ is not close to \mathbf{y} in terms of distance along the class manifold. Thus $\mathbf{x}_i^{(l)}$ and its tangent vectors should not be needed in the ℓ^1 -minimized approximation of \mathbf{y} .

A deeper notion of the parameter r is to view it as a rough estimate of the local neighborhood radius of the data set. More precisely, r estimates the distance from a sample within which its class manifold can be well-approximated by a tangent hyperplane (at that sample). Given X_{tr} and n , r is automatically computed, as described in Algorithm 2. In words, we set r to be the median distance between each training sample and its $(n + 1)$ st nearest neighbor (in the same class), where n , the number of neighbors in local PCA, is used to implicitly define the local neighborhood. It follows that r is a robust estimate of the local neighborhood radius, as learned from the training data.

This also explains our choice for the tangent vector scaling factor $c = r\gamma$ (in Step 12 of Algorithm 2), where $\gamma \sim \text{unif}(0, 1)$. Multiplying each tangent hyperplane basis vector $\mathbf{u}_j^{(l,i)}$, $1 \leq j \leq d$, by this scalar and then shifting it by its corresponding training sample $\mathbf{x}_i^{(l)}$ helps to ensure that the resulting tangent vector, included in the dictionary $D_{\mathbf{y}}$ if $\pm \mathbf{x}_i^{(l)}$ is sufficiently close to \mathbf{y} , lies in the local neighborhood of $\mathbf{x}_i^{(l)}$ on the l th class manifold.

Remark 3. If the test sample \mathbf{y} is far from the training data, defining r as in Algorithm 2 may produce

$D_{\mathbf{y}} = \emptyset$, i.e., there may be no training samples within that distance of \mathbf{y} . Thus to prevent this degenerate case, we use a slightly modified technique for setting r in practice. After assigning the median neighborhood radius $r_1 := \text{median}\{r_i^{(l)} \mid 1 \leq i \leq N_l, 1 \leq l \leq L\}$, we define r_2 to be the distance between the test sample \mathbf{y} and the closest training sample (up to sign). We then define the pruning parameter $r := \max\{r_1, r_2\}$. In the (degenerate) case that $r = r_2$, the dictionary consists of the closest training sample and its tangent vectors, leading to nearest neighbor classification instead of an algorithm error. However, experimental results indicate that the pruning parameter r is almost always equal to the median neighborhood radius r_1 , and so we leave this “technicality” out of the official algorithm statement to make it easier to interpret.

4.4. Computational Complexity and Storage Requirements

4.4.1. Computational complexity of SRC

When the ℓ^1 -minimization algorithm HOMOTOPY [31] is used, it is easy to see that the computational complexity of SRC is dominated by this step. This complexity is $O(N_{\text{tr}}m\kappa + m^2\kappa)$, where κ is the number of HOMOTOPY iterations [32]. HOMOTOPY has been shown to be relatively fast and good for use in robust face recognition [32]. In our experiments, we use it in all classification methods requiring ℓ^1 -minimization.

4.4.2. Computational complexity of LPCA-SRC

The computational complexity of the offline phase in LPCA-SRC (Algorithm 2) is

$$O\left(m \sum_{l=1}^L N_l^2 + N_{\text{tr}}mn\right), \quad (7)$$

whereas that of the online phase (Algorithm 3) is

$$O\left(N_{\text{tr}}m + \frac{N_{\mathbf{y}}}{d} \log\left(\frac{N_{\mathbf{y}}}{d}\right) + N_{\mathbf{y}}m\kappa + m^2\kappa\right). \quad (8)$$

Recall that $N_{\mathbf{y}}$ denotes the number of columns in the pruned dictionary $D_{\mathbf{y}}$. We note that the offline cost in Eq. (7) is based on the linear nearest neighbor search algorithm for simplicity; in practice there are faster methods. In our experiments, we used ATRIA (*Advanced Triangle Inequality Algorithm* [33]) via the MATLAB TSTOOL functions `nn_prepare` and `nn_search` [34]. The first function prepares the set of class l training samples $\mathcal{X}^{(l)}$ for nearest neighbor search at the onset, with the intention that subsequent runs of `nn_search` on this set are faster than simply doing a search without the preparation function. Other fast nearest neighbor search algorithms are available, for example, *k-d tree* [35]. The cost complexity estimates of these fast nearest neighbor search algorithms are somewhat complicated, and so we do not use them in Eq. (7). Hence, Eq. (7) could be viewed as the worst-case scenario.

Offline and online phases combined, the very worst-case computational complexity of LPCA-SRC is $O(N_{\text{tr}}^4)$, which occurs when the second-to-last term in Eq. (8) dominates: i.e., when (i) $N_{\mathbf{y}} \approx (d+1)N_{\text{tr}}$ (no pruning); (ii) $m \approx N_{\text{tr}}$ (large relative sample dimension); (iii) very large class manifold dimension estimate

d , so that d is relatively close to N_{tr} (note that this requires very large N_l for $1 \leq l \leq L$ by Eq. (6), which implies that L has to be very small); and (iv) $\kappa \approx m$ (many HOMOTOPY iterations). For small κ and N_l , $1 \leq l \leq L$, and when the pruning parameter r results in small $N_{\mathbf{y}}$ relative to N_{tr} , then the computational complexity reduces to approximately $O(N_{\text{tr}}m)$.

4.4.3. Storage requirements

The primary difference between the storage requirements for LPCA-SRC and SRC is that the offline phase of LPCA-SRC requires storing the matrix $D \in \mathbb{R}^{m \times (d+1)N_{\text{tr}}}$, which has a factor of $d+1$ as many columns as the matrix of training samples $X_{\text{tr}} \in \mathbb{R}^{m \times N_{\text{tr}}}$ stored in SRC. Hence the storage requirements of LPCA-SRC are at worst $(d+1)$ times the amount of storage required by SRC.

Though this potentially is a large increase, consider that in applications such as face recognition, it is expected that the intrinsic class manifold dimension be small, e.g., 3-5 [36]. Second, as we discussed in Remark 1 in Section 4.3.1, it is often sufficient to take d smaller than the actual intrinsic dimension (e.g., $d \in \{1, 2\}$) in LPCA-SRC. This, combined with the assumption that the original training set in SRC is not too large (so that the ℓ^1 -minimization problem in SRC can be solved fairly efficiently), suggests that the additional storage requirements of LPCA-SRC over SRC may not deter from the use of LPCA-SRC. We discuss this further with respect to our experimental results in Section 5.

5. Experiments

We tested the proposed classification algorithm on one synthetic database and three popular face databases. For all data sets, we used HOMOTOPY to solve the regularized versions of the ℓ^1 -minimization problems, i.e., Eq. (3) for SRC and Eq. (5) for LPCA-SRC, using version 2.0 of the L1 Homotopy toolbox [37].

5.1. Algorithms Compared

We compared LPCA-SRC to the original SRC, $\text{SRC}_{\text{pruned}}$ (a modification of SRC which we explain shortly), two versions of *tangent distance classification* (our implementations are inspired by Yang et al. [16]), *locality-sensitive dictionary learning SRC* [21], *k-nearest neighbors classification*, and *k-nearest neighbors classification over extended dictionary*.

- $\text{SRC}_{\text{pruned}}$: To test the efficacy of the tangent vectors in the LPCA-SRC dictionary, this modification of SRC prunes the dictionary of original training samples using the pruning parameter r , as in LPCA-SRC. $\text{SRC}_{\text{pruned}}$ is exactly LPCA-SRC without the addition of tangent vectors.
- *Tangent distance classification* (TDC1 and TDC2): We compared LPCA-SRC to two versions of tangent distance classification to test the importance of our algorithm's sparse representation framework. Both of our implementations begin by first finding a pruned matrix $D_{\mathbf{y}}^{\text{TDC}}$ that is very similar to the dictionary $D_{\mathbf{y}}$ in LPCA-SRC. In particular, $D_{\mathbf{y}}^{\text{TDC}}$ can be found using Algorithm 2 and Steps 1-10 in

Algorithm 3, *omitting Step 2 in each algorithm*. That is, neither the training nor test samples are ℓ^2 -normalized in the TDC methods; compared to the SRC algorithms, TDC1 and TDC2 are not sensitive to the energy of the samples. We emphasize that the resulting matrix $D_{\mathbf{y}}^{\text{TDC}}$ contains training samples that are nearby \mathbf{y} , as well as their corresponding tangent vectors.

In TDC1, we then divide $D_{\mathbf{y}}^{\text{TDC}}$ into the “subdictionaries” $D_{\mathbf{y}}^{(l)}$, where $D_{\mathbf{y}}^{(l)}$ contains the portion of $D_{\mathbf{y}}^{\text{TDC}}$ corresponding to class l . The test sample \mathbf{y} is next projected onto the space spanned by the columns of $D_{\mathbf{y}}^{(l)}$ to produce the vector $\hat{\mathbf{y}}^{(l)}$, and the final classification is performed using

$$\text{class_label}(\mathbf{y}) = \arg \min_{1 \leq l \leq L} \|\mathbf{y} - \hat{\mathbf{y}}^{(l)}\|_2.$$

Our second implementation, TDC2, is similar. Instead of dividing $D_{\mathbf{y}}^{\text{TDC}}$ according to class, however, we split it up according to training sample, obtaining the subdictionaries $D_{\mathbf{y}}^{(l,i)}$, where $D_{\mathbf{y}}^{(l,i)}$ contains the original training sample $\mathbf{x}_i^{(l)}$ and its tangent vectors. It follows that each subdictionary in TDC2 has $d+1$ columns. The given test sample \mathbf{y} is next projected onto the space spanned by the columns of $D_{\mathbf{y}}^{(l,i)}$ to produce $\hat{\mathbf{y}}_i^{(l)}$, a vector on the (approximate) tangent hyperplane at $\mathbf{x}_i^{(l)}$. The final classification is performed using

$$\text{class_label}(\mathbf{y}) = \arg \min_{1 \leq l \leq L} \left\{ \min_{1 \leq i \leq N_l} \|\mathbf{y} - \hat{\mathbf{y}}_i^{(l)}\|_2 \right\}.$$

- *Locality-sensitive dictionary learning SRC* (LSDL-SRC): Instead of directly minimizing the ℓ^1 -norm of the coefficient vector, LSDL-SRC replaces the regularization term in Eq. (3) of SRC with a term that forces large coefficients to occur only at dictionary elements that are close (in terms of an exponential distance function) to the given test sample. LSDL-SRC also includes a separate dictionary learning phase in which columns of the dictionary are selected from the columns of X_{tr} . We note that though the name “LSDL-SRC” contains the term “SRC,” this algorithm is less related to SRC than our proposed algorithm, LPCA-SRC. See Wei et al.’s paper [21] for their reasoning behind this name choice. However, the two algorithms do have very similar objectives, and we thought it important to compare LPCA-SRC and LSDL-SRC in order to validate our alternative approach.
- *k-nearest neighbors classification* (k NN): The test sample is classified to the most-represented class from among the nearest (in terms of Euclidean distance) k training samples (k is odd).
- *k-nearest neighbors classification over extended dictionary* (k NN-Ext): This is k NN over the columns of the (full) extended dictionary that includes the original training samples and their tangent vectors. Samples are not normalized at any stage.

5.2. Setting of Parameters

For the synthetic database, we used cross-validation at each instantiation of the training set to choose the best parameters n , λ , and d in LPCA-SRC. (Though the true class manifold dimension is known on

this database, we cannot always assume that this is the case.) We optimized the parameters consecutively as described in Section 4.3.2, each over its own set of discrete values according to our suggested guidelines and using Λ_{vals} as given in the example in Algorithm 5. We used the same approach for the parameter λ in SRC, the parameters n and λ in $\text{SRC}_{\text{pruned}}$, and the parameters n and d in the TDC algorithms. Finally, we used a similar procedure for the multiple parameters in LSDL-SRC (including its number of dictionary elements), and we also set k in $k\text{NN}$ and $k\text{NN-Ext}$ using cross-validation.

Our approach for the face databases was very similar, though in order to save computational costs, we set some parameter values according to previously published works. In particular, we set $\lambda = 0.001$ in LPCA-SRC, SRC, and $\text{SRC}_{\text{pruned}}$, as was used in SRC by Waqas et al. [20]. Additionally, we set most of the parameters in LSDL-SRC to the values used by its authors [21] on the same face databases, though we again used cross-validation to determine its number of dictionary elements.

5.3. Synthetic Database

This subsection is organized into two parts: We describe the synthetic database in Section 5.3.1, and we present our experimental findings in Section 5.3.2. Figures 2 and 3 and Table 2 show the accuracy and runtime results (as well as related information) respectively, for different versions of the synthetic database. A thorough discussion follows. Note that some algorithms from Section 5.1 (“Algorithms Compared”) have been excluded from these reported findings because of their poor performance, as we explain towards the end of Section 5.3.2. Finally, we briefly discuss the storage differences between LPCA-SRC and SRC and then summarize our results on the synthetic database.

5.3.1. Database description

The following synthetic database is easily visualized, and its class manifolds are nonlinear (though well-approximated by local tangent planes) with many intersections. Thus it is ideal for empirically comparing LPCA-SRC and SRC. The class manifolds are sinusoidal waves normalized to lie on S^2 , with underlying equations given by

$$\begin{aligned}x(t) &= \cos(t + \phi), \\y(t) &= \sin(t + \phi), \\z(t) &= A \sin(\omega t).\end{aligned}$$

We set $\omega = 3$ and $A = 0.5$, and we varied ϕ to obtain L classes. In particular, we set $\phi = 2\pi/(3l)$ for data in class $1 \leq l \leq L = 4$. For each training and test set, we generated the same number $N_0 = N_l$, $l = 1, \dots, L$, of samples in each class by (i) regularly sampling $t \in [0, 2\pi)$ to obtain the points $\mathbf{p}(t) = [x(t), y(t), z(t)]^T$; (ii) computing the normalized points $\mathbf{p}(t)/\|\mathbf{p}(t)\|_2$; (iii) appending 50 “noise dimensions” to obtain vectors in \mathbb{R}^{53} ; (iv) adding independent random noise to each coordinate of each point as drawn from the Gaussian

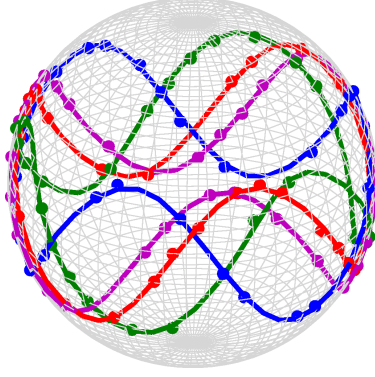


Figure 1: A realization of the first three coordinates of the synthetic database training set with $N_0 = 25$ and $\eta = 0.01$. Nodes denote training samples; colors denote classes. (See the color online copy.)

distribution $\mathcal{N}(0, \eta^2)$; and lastly (v) re-normalizing each point to obtain vectors of length $m = 53$ lying on S^{m-1} . We performed classification on the resulting data samples. Note that the reason why we turned the original \mathbb{R}^3 problem into a problem in \mathbb{R}^{53} was because SRC is designed for high-dimensional classification problems [2] and to make the problem more challenging. We emphasize that we did not apply any method of dimension reduction to this database.

Figure 1 shows the first three coordinates of a realization of the training set of the synthetic database. Note that the class manifold dimension is the same for each class and equal to 1. The signal-to-noise (SNR) ratios are displayed in Table 1 for $N_0 = 25$ and various values of noise level η . These results were obtained by averaging the mean training sample SNR over 100 realizations of the data set.

$\eta = 0.0001$	$\eta = 0.001$	$\eta = 0.005$	$\eta = 0.01$	$\eta = 0.015$	$\eta = 0.02$	$\eta = 0.03$	$\eta = 0.05$
62.85	42.84	28.86	22.86	19.35	16.89	13.45	9.25

Table 1: Mean training sample signal-to-noise ratio (in decibels) over 100 realizations of the synthetic database with $N_0 = 25$ and various values of noise level η .

5.3.2. Experimental results

We performed experiments on this database, first varying the number of training samples in each class and then varying the amount of noise. The results are presented in Figures 2 and 3 and Table 2; a discussion follows.

Accuracy results for varying class size. Figure 2 shows the average classification accuracy (over 100 trials) of the competitive algorithms as we varied the number of training samples in each class. We fixed the noise level $\eta = 0.001$. LPCA-SRC generally had the highest accuracy. On average, LPCA-SRC outperformed SRC by 3.5%, though this advantage slightly decreased as the sampling density increased and the tangent

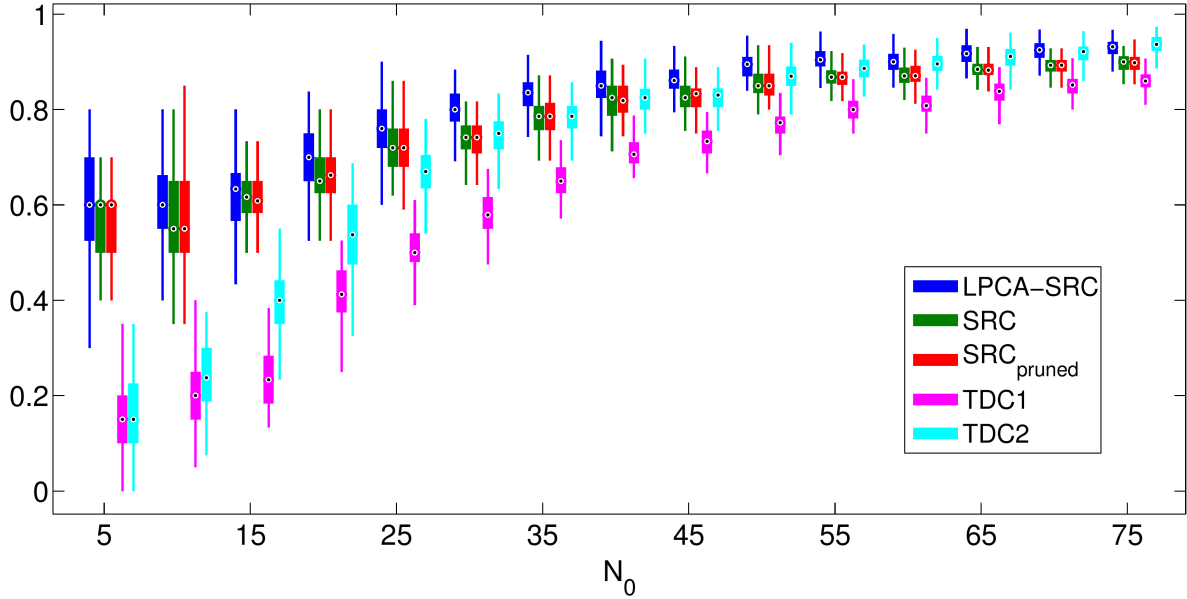


Figure 2: Box plot of the average classification accuracy (over 100 trials) of competitive algorithms on the synthetic database with varying training class size N_0 . We fixed $\eta = 0.001$. (See the color online copy.)

	$N_0 = 5$			$N_0 = 25$			$N_0 = 45$			$N_0 = 65$		
Algorithm	t	N	κ	t	N	κ	t	N	κ	t	N	κ
LPCA-SRC	11.2	56	2	68.8	80	3	115.3	42	3	159.2	30	2
SRC	4.5	20	2	39.9	100	3	104.6	180	3	162.8	260	3
SRC _{pruned}	7.1	20	2	54.1	79	3	130.2	146	3	206.0	201	3
TDC1	10.8	9	N/A	43.6	6	N/A	71.1	5	N/A	92.3	3	N/A
TDC2	19.5	3	N/A	57.0	2	N/A	93.4	2	N/A	125.4	2	N/A

Table 2: Average runtime in milliseconds (t), dictionary size (N), and number of HOMOTOPY iterations (κ) over 100 trials on the synthetic database with varying training class size N_0 . We fixed $\eta = 0.001$.

vectors became less useful, in the sense that there were often already enough nearby training samples in the ground truth class of \mathbf{y} to accurately approximate it without the addition of tangent vectors. SRC and SRC_{pruned} had comparable accuracy for all tried values of N_0 , indicating that the pruning parameter r was effective in removing unnecessary training samples from the SRC dictionary. Further, the increased accuracy of LPCA-SRC over SRC_{pruned} suggests that the tangent vectors in LPCA-SRC contributed meaningful class information.

The TDC methods performed relatively poorly for small values of N_0 . At low sampling densities, the TDC subdictionaries were poor models of the (local) class manifolds, leading to approximations of \mathbf{y} that were

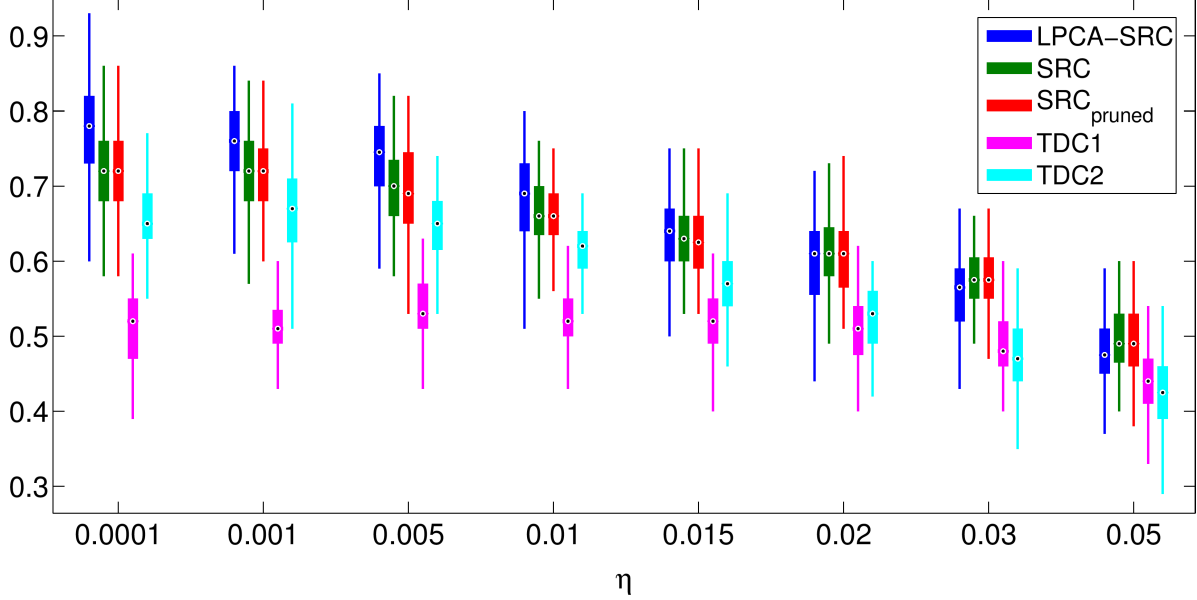


Figure 3: Box plot of the average classification accuracy (over 100 trials) of competitive algorithms on the synthetic databases with varying noise level η . We fixed $N_0 = 25$. (See the color online copy.)

often indistinguishable from each other and resulting in poor classification. Both TDC methods improved significantly as N_0 increased, with TDC2 outperforming TDC1 and in fact becoming comparable to LPCA-SRC for $N_0 \geq 60$. We attribute this to the extremely local nature of TDC2: It considers a single local patch on a class manifold at a time, rather than each class as a whole. Hence under dense sampling conditions, TDC2 effectively mimicked the successful use of sparsity in LPCA-SRC.

Accuracy results for varying noise. Figure 3 shows the average classification accuracy (over 100 trials) of the competitive algorithms as we varied the amount of noise. We fixed $N_0 = 25$. LPCA-SRC had the highest classification accuracy for low values of η (equivalently, when the SNR was high), outperforming SRC by as much as nearly 4%. For $\eta \geq 0.015$ (i.e., when the SNR dropped below 20 decibels), LPCA-SRC lost its advantage over SRC and SRC_{pruned}. This is likely due to noise degrading the accuracy of the tangent vectors. SRC and SRC_{pruned} had nearly identical accuracy for all values of η ; again, this illustrates that faraway training samples (as defined by the pruning parameter r) did not contribute to the ℓ^1 -minimized approximation of the test sample, and the increased accuracy of LPCA-SRC over SRC_{pruned} for low noise values demonstrates the efficacy of the tangent vectors in LPCA-SRC in these cases. We briefly note that when we vary the noise level for larger values of N_0 , the accuracy of the tangent vectors generally improves. As a result, we see that LPCA-SRC can tolerate higher values of η before being outperformed by SRC and SRC_{pruned}.

TDC2 outperformed TDC1 for all but the largest values of η , though both algorithms were outperformed

by the three SRC methods at this relatively low sampling density for the reasons discussed previously. For $\eta \geq 0.03$, TDC2 began performing worse than TDC1. We expect that the local patches represented by the subdictionaries in TDC2 became poor estimates of the (tangent hyperplanes of the) class manifolds as the noise increased, resulting in a decrease in classification accuracy.

Runtime results for varying class size. In Table 2, we display the runtime-related information of the competitive algorithms with varying training class size. (We do not show the runtime results for the case of varying noise; the results for varying class size are much more revealing.) In particular, we report the average runtime (in milliseconds), the number of columns in each algorithm’s dictionary (we refer to this as the “size” of the dictionary, as the sample dimension is fixed), and the number of HOMOTOPY iterations. These latter variables are denoted N and κ , respectively. The runtime does not include the time it took to perform cross-validation and is the total time (averaged over 100 trials) of performing classification on the entire database. In the case that the algorithm has separate offline and online phases (e.g., LPCA-SRC), both phases are included in this total. For the TDC methods, we report the average subdictionary sizes, and for conciseness, we display the results for only a handful of the values of N_0 . We use “N/A” to indicate that a particular statistic is not applicable to the given algorithm.

The dictionary sizes of LPCA-SRC, SRC, and SRC_{pruned} are quite informative. Recall that LPCA-SRC outperformed SRC and SRC_{pruned} (by more than 3%) for the shown values of N_0 . For $N_0 = 5$, the dictionary in LPCA-SRC was larger than that of the two other methods, adaptively retaining more samples to counter-balance the low sampling density. At large values of N_0 , LPCA-SRC took full advantage of the increased sampling density, stringently pruning the set of training samples and keeping only those very close to \mathbf{y} . Due to the resulting small dictionary, it had comparable runtime to SRC despite its additional cost of computing tangent vectors. In contrast, without the addition of tangent vectors, SRC_{pruned} was forced to keep a large number of training samples in its dictionary; the cost of the dictionary pruning step resulted in SRC_{pruned} running slower than SRC, despite its slightly smaller dictionary. (We note that one might expect that SRC_{pruned} would always have a smaller dictionary than LPCA-SRC since it does not include tangent vectors; this is not the case, as the value of the number-of-neighbors parameter n , and hence the pruning parameter r , may be different for the two algorithms.)

The TDC methods ran relatively fast, especially for large values of N_0 . This is expected, as these algorithms do not require ℓ^1 -minimization.

Reason for including only some of the algorithms discussed in Section 5.1. The algorithms LPCA-SRC, SRC, SRC_{pruned}, and the TDC methods significantly exceeded LSDL-SRC and the k NN methods in terms of accuracy in these experiments. In particular, these latter three methods were always outperformed by LPCA-SRC by at least 10% and often by as much as 30% – 40%. Though k NN-Ext generally performed better than k NN, neither method was competitive due to its inability to distinguish individual class manifolds

near intersections, a result of considering the classes in terms of a single sample (or tangent vector) at a time. On the other hand, LSDL-SRC was not *local enough*; despite its explicit locality term, this method was unable to distinguish the individual classes from within a local neighborhood of the test sample. Because of their poor performance, we do not report the results of these algorithms.

In contrast, the approximations in LPCA-SRC, SRC, and SRC_{pruned} typically contained nonzero coefficients solely at one or two dictionary elements bordering the test sample (up to sign) on the correct class manifold. That is, these approximations were *very* sparse, and this sparsity often resulted in correct classification. The TDC methods, though generally not as competitive as these first three algorithms, also showed relatively good performance; when there was a large enough number of training samples in each class, the TDC class-specific subdictionaries were effective in discriminating between classes.

Storage comparison. Though not reported in the above figures and table, the value of d used by LPCA-SRC (determined using cross-validation) was consistently in $\{1, 2\}$ (though larger values were tried). The median value of d used in LPCA-SRC in all experiments on the synthetic database was 1.3. Thus the storage required by LPCA-SRC was often twice that of SRC, per Section 4.4.3.

Summary. The experimental results on the synthetic database show that LPCA-SRC can achieve higher classification accuracy than SRC and similar methods when the class manifolds are sparsely sampled and the SNR is large. In these cases, the tangent vectors in LPCA-SRC help to “fill out” portions of the class manifolds that lack training samples. When the sampling density was sufficiently high, however, we saw that the tangent vectors in LPCA-SRC were less needed to provide an accurate, local approximation of the test sample, and thus LPCA-SRC offered a smaller advantage over SRC and SRC_{pruned}. Additionally, for higher noise (i.e., low SNR) cases, the computed tangent vectors were less reliable and the classification performance consequently deteriorated. With regard to runtime, LPCA-SRC appeared to adapt to the sampling density of the synthetic database, and though the addition of tangent vectors initially increased the dictionary size in LPCA-SRC, the online dictionary pruning step allowed for runtime comparable to SRC when the class sizes were large. The storage requirements of LPCA-SRC were often not more than twice those of SRC.

5.4. Face Databases

This subsection is organized as follows:

- We first explain our experimental setup. We describe the different face databases and state the training set sizes in Section 5.4.1, and in Sections 5.4.2 and 5.4.3, we describe the method of dimension reduction used on the raw samples and our approach to handling data samples with occlusion, respectively.
- We separate our classification results into two parts: Section 5.4.4 contains our results on the AR face database, and Section 5.4.5 contains our results on the Extended Yale B and ORL face databases. More precisely, Tables 3 and 4 contain the accuracy and runtime results for two versions of the AR face database; Tables 5, 6 and 7 show the same results for Extended Yale B and ORL. Again, these

databases are described in Section 5.4.1. The tables in each section are followed by a discussion of their results, as well as a comparison of the storage requirements between LPCA-SRC and SRC.

- In Section 5.4.6, we offer evidence to support our claim that the tangent vectors in LPCA-SRC can recover discriminative information lost during PCA transforms to low dimensions. We display the PCA-recovered tangent vectors and compare them to the original samples (without PCA transform) as well as the recovered samples (after PCA transform).
- Lastly, Section 5.4.7 contains a summary of our experimental findings on the face databases.

5.4.1. Database description

The *AR Face Database* [13] contains 70 male and 56 female subjects photographed in two separate sessions held on different days. Each session produced 13 images of each subject, the first seven with varying lighting conditions and expressions, and the remaining six images occluded by either sunglasses or scarves under varying lighting conditions. Images were cropped to 165×120 pixels and converted to grayscale. In our experiments, we selected the first 50 male subjects and first 50 female subjects, as was done in several papers (e.g., Wright et al. [2]), for a total of 100 classes. We performed classification on two versions of this database. The first, which we call “AR-1,” contains the 1400 un-occluded images from both sessions. The second version, “AR-2,” consists of the images in AR-1 as well as the 600 occluded images (sunglasses and scarves) from Session 1.

The *Extended Yale Face Database B* [38] contains 38 classes (subjects) with about 64 images per class. The subjects were photographed from the front under various lighting conditions. We used the version of Extended Yale B that contains manually-aligned, cropped, and resized images of dimension 192×168 .

The *Database of Faces* (formerly “The ORL Database of Faces”) [39] contains 40 classes (subjects) with 10 images per class. The subjects were photographed from the front against dark, homogeneous backgrounds. The sets of images of some subjects contain varying lighting conditions, expressions, and facial details. Each image in ORL is initially of 92×112 pixels.

Given existing work on the manifold structure of face databases (e.g., that of Saul and Roweis [11], He et al. [12], and Lee et al. [36]), we make the following suppositions: Since images in each class in AR-1 and AR-2 have extreme variations in lighting conditions and differing expressions, the class manifolds of these databases may be nonlinear. Further, the natural occlusions contained in AR-2 make these class manifolds *highly* nonlinear. Alternatively, since the images in each class in Extended Yale B differ primarily in lighting conditions, the class manifolds may be nearly linear. Lastly, since the images in some classes in ORL differ in both lighting conditions and expression, these class manifolds may be nonlinear; however, since the variations are small, these manifolds may be well-approximated by linear subspaces.

With regard to sampling density, we reiterate that Extended Yale B has large class sizes compared to AR and ORL. In our experiments, we randomly selected the same number of samples in each class to use

for training, i.e., we set $N_0 \equiv N_l$, $1 \leq l \leq L$, where N_0 was half the number of samples in each class.² We used the remaining samples for testing.

5.4.2. Dimension reduction

To perform dimension reduction on the face databases, we used (global) PCA to transform the raw images to $m_{\text{PCA}} \in \{30, 56, 120\}$ dimensions before performing classification. Similar values for m_{PCA} were used by Wright et al. [2]. For the remainder of this paper, we will refer to the PCA-compressed versions of the raw face images as “feature vectors” and m_{PCA} as the “feature dimension.” We note that the data was not centered (around the origin) in the PCA transform space.

5.4.3. Handling occlusion

Since AR-2 contains images with occlusion, we considered using the “occlusion version” of SRC (with analogous modifications to LPCA-SRC and $\text{SRC}_{\text{pruned}}$) on this database. As discussed by Wright et al. [2], this model assumes that \mathbf{y} is the summation of the (unknown) true test sample \mathbf{y}_0 and an (unknown) sparse error vector. The resulting modified ℓ^1 -minimization problem consists of appending the dictionary of training samples with the identity matrix $I \in \mathbb{R}^{m \times m}$ and decomposing \mathbf{y} over this augmented dictionary. For more details, see Section 3.2 of the SRC paper [2].

However, the context in which Wright et al. use the occlusion version of SRC on the AR database is critically different than our experimental setup here [2]. In the SRC paper, the samples with occlusion make up the test set. In our case, both the training and test set contain samples with and without occlusion. As a consequence, occluded samples in the training set can be used to express test samples with occlusion, and on the other hand, the use of the identity matrix to extend the dictionary in SRC results in too much error allowed in the approximation of un-occluded samples. Correspondingly, we see much worse classification performance in SRC when we use its occlusion version on AR-2. Hence, we stick to Algorithm 1 (the original version of SRC) on all face databases.

5.4.4. AR Face Database results

Accuracy results on AR. Table 3 displays the average accuracy and standard error over 10 trials for the two versions of AR. LPCA-SRC had substantially higher classification accuracy than the other methods on both versions of AR with $m_{\text{PCA}} = 30$. This suggests that the tangent vectors in LPCA-SRC were able to recover important class information lost in the stringent PCA dimension reduction. As m_{PCA} increased, however, the methods SRC, $\text{SRC}_{\text{pruned}}$, and LSDL-SRC became more competitive, as more discriminative information was retained in the feature vectors and less needed to be provided by the LPCA-SRC tangent vectors. $\text{SRC}_{\text{pruned}}$ had comparable accuracy to SRC, indicating that, once again, training samples could be

²Since the class sizes vary slightly in Extended Yale B, we set $N_0 = 32$ on this database.

	AR-1						AR-2					
	$m_{\text{PCA}} = 30$		$m_{\text{PCA}} = 56$		$m_{\text{PCA}} = 120$		$m_{\text{PCA}} = 30$		$m_{\text{PCA}} = 56$		$m_{\text{PCA}} = 120$	
Algorithm	Acc	SE	Acc	SE	Acc	SE	Acc	SE	Acc	SE	Acc	SE
LPCA-SRC	0.8663	4.1	0.9544	2.3	0.9711	1.7	0.7328	6.0	0.8844	3.6	0.9512	2.6
SRC	0.8273	4.2	0.9357	2.6	0.9631	1.6	0.6945	4.3	0.8713	2.0	0.9450	2.4
SRC _{pruned}	0.8277	4.8	0.9353	3.8	0.9651	1.8	0.7092	4.0	0.8781	2.7	0.9459	2.5
TDC1	0.8046	6.5	0.8430	5.1	0.8634	5.9	0.6899	4.1	0.7603	3.7	0.7985	4.5
TDC2	0.7549	19.4	0.8137	11.9	0.8303	15.4	0.6422	16.1	0.7386	3.3	0.7735	4.4
LSDL-SRC	0.8184	4.0	0.9424	2.0	0.9756	0.9	0.6585	5.6	0.8610	2.2	0.9498	2.9
k NN	0.5846	4.6	0.6301	8.0	0.6461	4.9	0.4100	3.0	0.4297	5.0	0.4554	3.2
k NN-Ext	0.6036	4.5	0.6487	8.2	0.6677	4.7	0.4311	3.7	0.4526	2.9	0.4794	5.7

Table 3: Average accuracy and standard error ($\times 10^{-3}$) over 10 trials on AR.

	AR-1								
	$m_{\text{PCA}} = 30$			$m_{\text{PCA}} = 56$			$m_{\text{PCA}} = 120$		
Algorithm	t	N	κ	t	N	κ	t	N	κ
LPCA-SRC	7253	435	61	12496	676	87	19068	795	112
SRC	6114	700	51	8875	700	72	13574	700	99
SRC _{pruned}	3763	231	39	5099	226	49	6897	232	60
TDC1	11816	16	N/A	14239	16	N/A	24296	19	N/A
TDC2	8895	5	N/A	16786	5	N/A	36682	5	N/A
LSDL-SRC	7776	440	N/A	8552	470	N/A	9720	490	N/A
k NN	13	700	N/A	18	700	N/A	29	700	N/A
k NN-Ext	102	2170	N/A	132	2240	N/A	253	2660	N/A
	AR-2								
	$m_{\text{PCA}} = 30$			$m_{\text{PCA}} = 56$			$m_{\text{PCA}} = 120$		
Algorithm	t	N	κ	t	N	κ	t	N	κ
LPCA-SRC	10533	478	58	35269	1593	10	56169	1690	151
SRC	11394	1000	58	17674	1000	85	27743	1000	121
SRC _{pruned}	11118	788	54	16631	775	77	24880	767	107
TDC1	20557	25	N/A	27515	26	N/A	43073	26	N/A
TDC2	20930	6	N/A	47571	6	N/A	103796	6	N/A
LSDL-SRC	22698	750	N/A	16337	620	N/A	22191	710	N/A
k NN	15	1000	N/A	21	1000	N/A	37	1000	N/A
k NN-Ext	128	4300	N/A	152	3600	N/A	294	4400	N/A

Table 4: Average runtime in milliseconds (t), dictionary size (N), and number of HOMOTOPY iterations (κ) over 10 trials on AR.

removed from the SRC dictionary using the pruning parameter r without decreasing classification accuracy. In some cases, the removal of these faraway training samples slightly improved class discrimination.

For the most part, the other algorithms performed poorly on AR. The exception was LSDL-SRC, which had comparable accuracy to LPCA-SRC for $m_{\text{PCA}} = 120$ (slightly outperforming it for AR-1) and beat SRC on AR-1 for $m_{\text{PCA}} = 56$. However, LSDL-SRC had lower accuracy than the SRC algorithms for $m_{\text{PCA}} = 30$ on both versions of this database. In contrast, the TDC methods performed relatively better for $m_{\text{PCA}} = 30$ than for larger values of m_{PCA} due to their more effective use of tangent vectors at this small feature dimension. Overall, however, their class-specific dictionaries were not as effective on this nonlinear, sparsely sampled database as the multi-class dictionaries of the previously-discussed algorithms. Further, TDC2 often had notably high standard error, presumably because of its sensitivity to the value of the manifold dimension estimate d . This could perhaps be mitigated by using a different cross-validation procedure. Lastly, k NN and k NN-Ext had the lowest classification accuracies, though k NN-Ext offered a slight improvement over k NN. Both methods consistently selected $k = 1$ during cross-validation.

Runtime results on AR. Table 4 displays the average runtime and related results (over 10 trials) of the various classification algorithms for both versions of AR. Again, the runtime does not include the time it took to perform cross-validation and is the total time (averaged over 10 trials) of performing classification on the entire database (offline and online phases both included when applicable). The “dictionary size” N for k NN and k NN-Ext refers to the average size of the set from which the k -nearest neighbors are selected (e.g., for k NN, $N = N_{\text{tr}}$).

The generally large dictionary sizes of LPCA-SRC (and its consequently long runtimes) indicate that minimal dictionary pruning often occurred. Thus LPCA-SRC was generally slower than SRC and SRC_{pruned}. However, on AR-2 with $m_{\text{PCA}} = 30$, LPCA-SRC was able to eliminate many training samples from its dictionary, due to its effective use of tangent vectors on the (presumably) highly-nonlinear class manifolds of AR-2. At this low feature dimension, the computed tangent vectors contained more class discriminative information than nonlocal training samples, likely allowing for a more accurate—and local—approximation of \mathbf{y} on its ground truth class manifold. LPCA-SRC was faster than SRC and SRC_{pruned} (which kept a large number of training samples) in this case, and this is impressive, considering that LPCA-SRC also outperformed these methods by nearly 4% and more than 2%, respectively.

Despite not requiring ℓ^1 -minimization, the TDC methods were often the slowest algorithms on the AR databases. We suspect that this is largely due to the relatively large number of classes in AR—recall that both TDC methods must compute least squares solutions (in TDC2, sometimes many of them) for each class represented in the pruned dictionary $D_{\mathbf{y}}^{\text{TDC}}$. Further, TDC2 selected a relatively large value of d during cross-validation (presumably so that its subdictionaries would contain a wider “snapshot” of the class manifolds), which made it even less efficient. The runtime of LSDL-SRC, unlike those of most of the other

algorithms, was fairly insensitive to the feature dimension, and as a result, LSDL-SRC was relatively efficient for $m_{\text{PCA}} \in \{56, 120\}$. However, the expense of its dictionary learning phase for $m_{\text{PCA}} = 30$, at which the ℓ^1 -minimization algorithm in the SRC methods could be solved efficiently, resulted in LSDL-SRC’s relatively slow runtime. Both k NN methods ran significantly faster than all the other methods.

Storage comparison on AR. The value of d selected using cross-validation in LPCA-SRC on the two AR face databases was never larger than 3. The median value for d was 1.1. So the storage requirements of LPCA-SRC were often twice that of SRC, but occasionally 3-4 times as much. (Recall that LPCA-SRC requires $(d + 1)$ times the amount of storage as SRC, per Section 4.4.3.)

5.4.5. Extended Yale Face Database B and Database of Faces (“ORL”) results

	Extended Yale B						ORL					
	$m_{\text{PCA}} = 30$		$m_{\text{PCA}} = 56$		$m_{\text{PCA}} = 120$		$m_{\text{PCA}} = 30$		$m_{\text{PCA}} = 56$		$m_{\text{PCA}} = 120$	
Algorithm	Acc	SE	Acc	SE	Acc	SE	Acc	SE	Acc	SE	Acc	SE
LPCA-SRC	0.9049	2.9	0.9530	1.7	0.9710	1.6	0.9507	24.0	0.9600	18.0	0.9602	17.0
SRC	0.8803	2.6	0.9371	2.8	0.9633	1.4	0.9374	24.0	0.9437	22.9	0.9422	18.8
SRC _{pruned}	0.8804	2.7	0.9371	2.6	0.9635	1.5	0.9506	23.7	0.9580	23.5	0.9605	19.3
TDC1	0.8568	10.0	0.9285	2.0	0.9446	2.8	0.9364	27.1	0.9457	25.4	0.9455	21.1
TDC2	0.8826	3.9	0.9093	2.8	0.9283	3.5	0.9351	29.8	0.9429	31.1	0.9418	23.3
LSDL-SRC	0.7495	4.8	0.8774	2.5	0.9492	2.0	0.9358	25.2	0.9515	19.9	0.9251	24.3
k NN	0.4300	3.5	0.5346	2.6	0.6245	3.8	0.9332	26.3	0.9387	24.7	0.9396	23.2
k NN-Ext	0.5464	6.9	0.6321	5.6	0.7058	5.4	0.9338	30.5	0.9412	28.9	0.9386	23.9

Table 5: Average accuracy and standard error ($\times 10^{-4}$) on Extended Yale B (over 10 trials) and ORL (over 50 trials).

Accuracy results on Extended Yale B and ORL. Table 5 displays the average accuracy and standard error for Extended Yale B (over 10 trials) and ORL (over 50 trials). On Extended Yale B, LPCA-SRC had the highest accuracy for all m_{PCA} , though as we saw on the AR database, this advantage decreased as m_{PCA} increased and SRC became more competitive. SRC and SRC_{pruned} had very similar accuracy, indicating that training samples excluded from the dictionary via the pruning parameter r did not provide class information in the SRC framework. TDC1 and TDC2 had consistently mediocre performance, neither one outperforming the other over all settings of m_{PCA} , and LSDL-SRC improved as m_{PCA} increased, analogous to its behavior on AR. However, LSDL-SRC was clearly outperformed by LPCA-SRC, even for $m_{\text{PCA}} = 120$, suggesting that the improved approximations in LPCA-SRC via its use of tangent vectors were more effective (even at this high feature dimension) than the procedure in LSDL-SRC. Along these same lines, the tangent vectors in k NN-Ext offered a considerable improvement over k NN, though once again both methods reported lower

	$m_{\text{PCA}} = 30$			$m_{\text{PCA}} = 56$			$m_{\text{PCA}} = 120$		
Algorithm	t	N	κ	t	N	κ	t	N	κ
LPCA-SRC	29204	1922	75	72122	3359	120	141966	3785	182
SRC	15584	1216	62	24697	1216	91	41939	1216	137
SRC _{pruned}	15915	1111	61	23813	1112	88	40504	1115	131
TDC1	8098	20	N/A	27620	59	N/A	42828	59	N/A
TDC2	11675	6	N/A	23506	6	N/A	56006	6	N/A
LSDL-SRC	67295	1186	N/A	53031	1003	N/A	38731	821	N/A
k NN	17	1216	N/A	26	1216	N/A	49	1216	N/A
k NN-Ext	172	5350	N/A	251	4742	N/A	443	4864	N/A

Table 6: Average runtime in milliseconds (t), dictionary size (N), and number of HOMOTOPY iterations (κ) over 10 trials on Extended Yale B.

	$m_{\text{PCA}} = 30$			$m_{\text{PCA}} = 56$			$m_{\text{PCA}} = 120$		
Algorithm	t	N	κ	t	N	κ	t	N	κ
LPCA-SRC	539	59	26	730	72	34	1221	111	50
SRC	854	200	40	1337	200	57	2087	200	81
SRC _{pruned}	254	19	12	343	26	16	530	39	24
TDC1	121	1	N/A	162	1	N/A	344	1	N/A
TDC2	117	3	N/A	233	3	N/A	532	3	N/A
LSDL-SRC	1040	116	N/A	1088	121	N/A	931	102	N/A
k NN	8	200	N/A	8	200	N/A	9	200	N/A
k NN-Ext	25	568	N/A	28	592	N/A	38	568	N/A

Table 7: Average runtime in milliseconds (t), dictionary size (N), and number of HOMOTOPY iterations (κ) over 50 trials on ORL.

accuracy than all the other algorithms. As on AR, the k NN methods consistently selected $k = 1$ during cross-validation.

On ORL, LPCA-SRC and SRC_{pruned} had comparable accuracy and outperformed SRC. This indicates that: (i) the pruning parameter r in LPCA-SRC and SRC_{pruned} was *helpful* to classification (instead of simply being benign); and (ii) the tangent vectors computed in LPCA-SRC were not. With regard to (i), it must be the case that faraway training samples—those in different classes from the test sample—contributed significantly to the approximation of the test sample in SRC, negatively affecting classification performance. This is an example of *sparsity not necessarily leading to locality* (as it is relevant to class discrimination), as discussed in the LSDL-SRC paper [21]. With regard to (ii), we suspect that the tangent vectors in LPCA-

SRC were simply *unnneeded* to improve the classification performance on ORL. Though the approximations in SRC contained nonzero coefficients at training samples not in the same class as \mathbf{y} —presumably because of the sparse sampling and nonlinear structure of the class manifolds—many of these wrong-class training samples could be eliminated simply based on their distance to \mathbf{y} . This suggests that ORL’s class manifolds can be fairly well-separated via Euclidean distance. An additional reason for (ii) was because the PCA transform to the dimensions specified in this experiment did not result in a loss of too much information, at least compared to AR and Extended Yale B. See Table 8 at the end of Section 5.4.6 for this comparison.

All of the remaining methods performed relatively well on ORL. The accuracies of TDC1 and TDC2 were similar and comparable to those of SRC. We ascertained that the success of the TDC methods was not due to their use of tangent vectors but instead the result of their “per-class” approximations of the test sample. This approach was very effective on the (presumably) well-separated class manifolds of ORL. Strikingly, the accuracy of LSDL-SRC was relatively low for $m_{\text{PCA}} = 120$, opposite to the trend we saw on the previous face databases. The performance of LSDL-SRC could be improved for $m_{\text{PCA}} = 120$ on this database if the samples were centered (around the origin) after PCA dimension reduction. However, we confirmed that LDSL-SRC was still outperformed by LPCA-SRC in this case (albeit by a smaller margin), and its performance with centering on the other face databases was much worse than our reported results. In contrast to the results on Extended Yale B, $k\text{NN-Ext}$ only provided a slight increase in accuracy over $k\text{NN}$, with the tangent vectors mimicking their unnecessary role in LPCA-SRC on this database. The value $k = 1$ was consistently selected by both $k\text{NN}$ and $k\text{NN-Ext}$ during cross-validation.

Runtime results on Extended Yale B and ORL. Tables 6 and 7 show the runtime and related results for the Extended Yale B and ORL experiments, respectively. LPCA-SRC had much longer runtimes than SRC on Extended Yale B, especially as m_{PCA} increased. This was due to a combination of large values for d selected during cross-validation and the tangent vectors’ decreasing efficacy at larger feature dimensions. However, the dictionary pruning procedure in LPCA-SRC actually eliminated a large number of training samples for all m_{PCA} ; once again, the computed tangent vectors contained more class-discriminating information than the eliminated nonlocal training samples, especially at lower feature dimensions for which details provided by these tangent vectors were especially needed. The (presumed) linearity of the class manifolds of Extended Yale B, combined with this database’s relatively dense sampling, lent itself well to the accurate computation of tangent vectors—part of the reason why LPCA-SRC used so many of them. Viewing these points as newly-generated and nearby training samples, LPCA-SRC’s boost in accuracy over SRC can be viewed as an argument for locality in classification. We note that we might be able to decrease the value of d in LPCA-SRC while still maintaining an advantage over SRC (see the discussion in Section 4.3.1); our cross-validation procedure is designed to obtain the highest accuracy with no regard to computational cost.

On Extended Yale B, the TDC methods ran relatively more quickly (compared to the other algorithms) than on AR, presumably due to the much smaller number of classes on this database; both had runtimes typically between those of LPCA-SRC and SRC. Again, we see that LSDL-SRC had a relatively slow runtime for $m_{\text{PCA}} = 30$ and became more competitive as m_{PCA} increased. Though both $k\text{NN}$ and $k\text{NN-Ext}$ were very fast, the large “dictionary sizes” in $k\text{NN-Ext}$ made this algorithm clearly the slower of the two methods.

On ORL, LPCA-SRC and SRC had comparable runtimes, a result of rigorous dictionary pruning in LPCA-SRC. This algorithm and $\text{SRC}_{\text{pruned}}$ retained roughly the same number of training samples in their respective dictionaries, and the latter was notably fast, running in about half the time as SRC. The remaining algorithms were even more efficient. TDC1 and TDC2 had comparable runtimes, both running faster than LSDL-SRC. As before, $k\text{NN}$ and $k\text{NN-Ext}$ had the fastest runtimes; the former was faster than the latter.

Storage comparison on Extended Yale B and ORL. Since d was often large in LPCA-SRC on Extended Yale B, the algorithm’s storage requirements were generally 4-5 times that of SRC. However, as mentioned, performance accuracy might still be maintained if d were made smaller, thus decreasing the amount of storage required. On ORL, d in LPCA-SRC was selected by cross-validation to be 2 in nearly all cases (though $d = 1$ and $d = 3$ occurred rarely), and so the storage requirements of LPCA-SRC on this database were typically three times that of SRC.

5.4.6. *Tangent vectors and PCA feature dimension*

In this section, we offer evidence to support our claim that the tangent vectors in LPCA-SRC can recover discriminative information lost during PCA transforms to low dimensions. Thus LPCA-SRC can offer a clear advantage over SRC in these cases, as we saw in experimental results on AR and Extended Yale B.

In Figures 4-6, we display three versions of three example images from AR-1. The first version is the original image (before PCA dimension reduction), the second version is the recovered image from PCA dimension reduction to dimension $m_{\text{PCA}} = 30$, and the third version is the recovered corresponding tangent vector computed in LPCA-SRC. In each case, the tangent vector contains details of the original image not found in the recovered image, supporting our claim that the tangent vectors in LPCA-SRC can recover some (but not all) of the information lost in stringent PCA dimension reduction.

Towards quantifying what we mean by “stringent,” Table 8 lists the average energy (over 10 trials) retained in the first m_{PCA} left-singular vectors of the face database training sets, along with the percent improvement in the accuracy of LPCA-SRC with respect to that of SRC and $\text{SRC}_{\text{pruned}}$. We reiterate that the addition of tangent vectors did not increase classification accuracy on ORL. Taking this into account, we see a correlation between the efficacy of tangent vectors in LPCA-SRC and the stringency of the PCA dimension reduction.

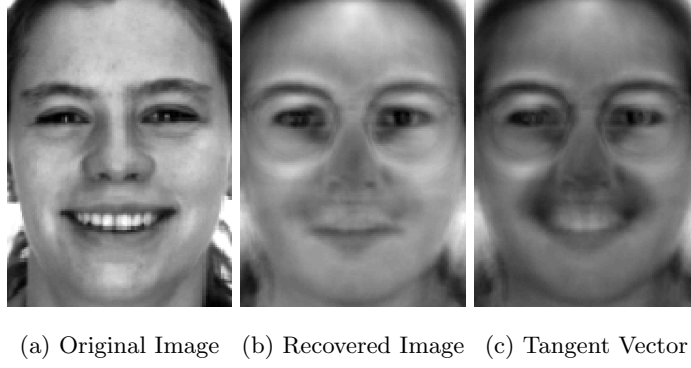


Figure 4: The tangent vector does a much better job of displaying facial details conveying “happiness” (displayed in the original image) than the recovered image. Images (b) and (c) were recovered from PCA dimension $m_{\text{PCA}} = 30$.

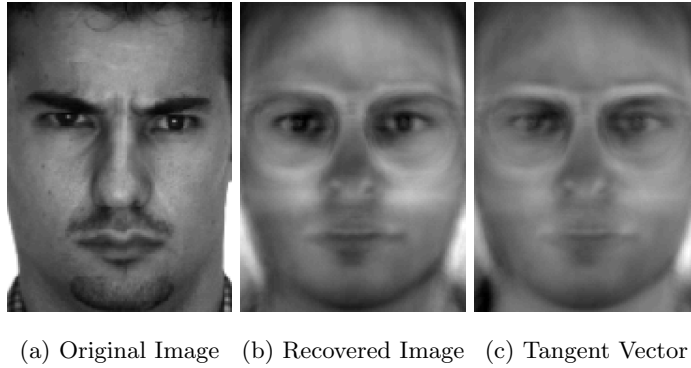


Figure 5: The tangent vector does a better job of displaying facial details conveying “anger” (displayed in the original image) than the recovered image, most notably in the subject’s eyes and eyebrows. Images (b) and (c) were recovered from PCA dimension $m_{\text{PCA}} = 30$.

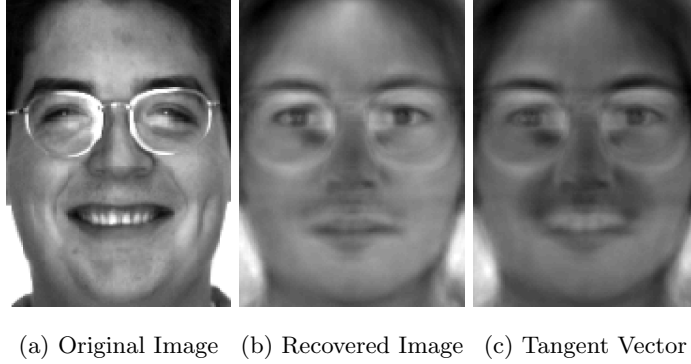


Figure 6: The tangent vector shows the subject’s smile, in particular, its shape, shading, and teeth and cheek detail (notice the right dimple), better than the recovered image. Images (b) and (c) were recovered from PCA dimension $m_{\text{PCA}} = 30$.

	$m_{\text{PCA}} = 30$		$m_{\text{PCA}} = 56$		$m_{\text{PCA}} = 120$	
		% Increased Acc.		% Increased Acc.		% Increased Acc.
Database	Energy	SRC / SRC _{pruned}	Energy	SRC / SRC _{pruned}	Energy	SRC / SRC _{pruned}
AR-1	0.4527	3.90/3.86	0.5322	1.87/1.91	0.6522	0.80/0.60
AR-2	0.4137	3.83/2.36	0.4884	1.31/0.63	0.5988	0.62/0.53
Extended Yale B	0.3954	2.46/2.45	0.4803	1.59/1.59	0.6055	0.77/0.74
ORL	0.5385	1.34/0.05	0.6581	1.26/-0.04	0.8487	1.73/0.03

Table 8: Average energy retained in PCA dimension reduction (over 10 trials) to various dimensions m_{PCA} on the face database training sets, as well as the average increase in classification accuracy of LPCA-SRC over SRC and SRC_{pruned}.

5.4.7. Summary

The experimental results on face databases show that LPCA-SRC can achieve higher accuracy than SRC in cases of low sampling and/or nonlinear class manifolds and small PCA feature dimension. We showed that LPCA-SRC had a significant advantage over SRC and the other algorithms for the small class sizes and nonlinear class manifolds of the AR database when the feature dimension was low. We also showed that LPCA-SRC could improve classification on Extended Yale B and ORL through its use of tangent vectors to provide a local approximation of the test sample and its discriminating pruning parameter, respectively.

The runtime of LPCA-SRC was sometimes much longer than that of SRC, although this was less often seen for small feature dimensions, at which LPCA-SRC tended to excel. The size of the dictionary in LPCA-SRC was observed to be a good predictor of the relationship between the runtimes of LPCA-SRC and SRC,

and this could easily be computed (given estimates of the parameters n and d) before deciding between the two methods. Though LPCA-SRC required no more than twice the memory of SRC on the AR database, its storage requirements were as much as 4-5 those of SRC on Extended Yale B. We acknowledge that using this much storage space may be undesirable. However, estimating d beforehand and possibly using a smaller value of d than that determined by cross-validation (e.g., $d \in \{1, 2\}$) may allow the algorithm to run within acceptable memory while still achieving a boost in accuracy over SRC.

To validate our claim that the tangent vectors in LPCA-SRC can contain information lost in stringent PCA dimension reduction, we provided examples from the AR database. We also compared the energy retained in PCA dimension reduction with the increase in accuracy in LPCA-SRC over SRC and saw that there was a correlation.

6. Further Discussion and Future Work

This paper presented a modification of SRC called *local principal component analysis SRC*, or “LPCA-SRC.” Through the use of tangent vectors, LPCA-SRC is designed to increase the sampling density of training sets and thus improve class discrimination on databases with sparsely sampled and/or nonlinear class manifolds. The LPCA-SRC algorithm computes basis vectors of approximate tangent hyperplanes at the training samples in each class and replaces the dictionary of training samples in SRC with a local dictionary (that is constructed based on each test sample) computed from shifted and scaled versions of these vectors and their corresponding training samples. Using a synthetic database and three face databases, we showed that LPCA-SRC can regularly achieve higher accuracy than SRC in cases of sparsely sampled and/or nonlinear class manifolds, low noise, and relatively small PCA feature dimension.

To address the issue of parameter setting, we recommended a consecutive parameter cross-validation procedure and gave detailed guidelines (including specific examples) for its use. We also briefly discussed alternative methods for determining the class manifold dimension estimate d . It is important to note that in the case of small training sets, e.g., many face recognition problems, there are few options for the number-of-neighbors parameter n —and consequently for d by Eq. (6)—and so these values can easily be set using cross-validation, as in our experiments. When the training sets are very small (i.e., $N_l = 4$ or 5), one could simply set n to its maximum value, i.e., $n = N_{l_{\min}} - 2$, per Eq. (6). On the other hand, simply setting $d = 1$ may suffice, especially when minimizing algorithm runtime and/or storage requirements is paramount.

One disadvantage of this method is its high computational cost and storage requirements. SRC is already expensive due to its ℓ^1 -minimization procedure; in LPCA-SRC, the computation of tangent vectors is added to the algorithm’s workload. The size of the dictionary in LPCA-SRC may be larger or smaller than that of SRC, depending on the LPCA-SRC parameters n and d and the effect of the pruning parameter r . Thus LPCA-SRC can be slower or faster than SRC. Further, the storage required by LPCA-SRC is $(d + 1)$ times

that of SRC, which may be prohibitive when d is large. As mentioned, simple computations based on the training set could render relative cost and storage estimates of using LPCA-SRC instead of SRC, and a smaller value of d than that found using cross-validation (e.g., $d = 1$) may be used successfully.

Additionally, as we saw on the synthetic database, the usefulness of the tangent vectors in LPCA-SRC decreases as the noise level in the training data increases. This problem could potentially be alleviated by using the method proposed by Kaslovsky and Meyer [40] to estimate clean points on the manifolds from noisy samples and then computing the tangent vectors at these points. Note that the case of large training sample noise was the only case for which we saw LPCA-SRC not obtain higher accuracy than SRC. Thus LPCA-SRC should be preferred over SRC in low noise scenarios on either small-scale problems (e.g., the size of ORL) or when achieving a modest (e.g., 1% – 4%) boost in accuracy is worth potentially higher computational cost.

Open questions regarding LPCA-SRC include whether or not the aforementioned general trends hold for different methods of dimension reduction besides PCA. Additionally, one could compare the performance of the “group” or “per-class” methods of the above representation-based algorithms, in which test samples are approximated using class-specific dictionaries (similarly to as in TDC1). Lastly, one could gain insight into the role of ℓ^1 -minimization in SRC by comparing LPCA-SRC and SRC_{pruned} to versions of these algorithms that replace the ℓ^1 -norm with the ℓ^2 -norm, analogous to the work of Zhang et al. in their *collaborative representation-based representation* model [41]. This is part of our ongoing work, which we hope to report at a later date.

Acknowledgments

C. Weaver’s research on this project was conducted with government support under contract FA9550-11-C-0028 and awarded by DoD, Air Force Office of Scientific Research, National Defense Science and Engineering Graduate (NDSEG) Fellowship, 32 CFR 168a. She was also supported by National Science Foundation VIGRE DMS-0636297 and NSF DMS-1418779. N. Saito was partially supported by ONR grants N00014-12-1-0177 and N00014-16-1-2255, as well as NSF DMS-1418779.

References

- [1] Y. LeCun, L. Bottou, Y. Bengio, P. Haffner, Gradient-based learning applied to document recognition, *Proceedings of the IEEE* 86 (11) (1998) 2278–2324. doi:10.1109/5.726791.
- [2] J. Wright, A. Y. Yang, A. Ganesh, S. S. Sastry, Y. Ma, Robust face recognition via sparse representation, *IEEE Trans. Pattern Anal. Mach. Intell.* 31 (2) (2009) 210–227. doi:10.1109/TPAMI.2008.79.

- [3] H. Cevikalp, H. S. Yavuz, M. A. Cay, A. Barkana, Two-dimensional subspace classifiers for face recognition, *Neurocomputing* 72 (46) (2009) 1111 – 1120. doi:10.1016/j.neucom.2008.02.015.
- [4] R. Patel, N. Rathod, A. Shah, Comparative analysis of face recognition approaches: A survey, *International Journal of Computer Applications* 57 (17) (2012) 50–69.
- [5] X. Tan, S. Chen, Z.-H. Zhou, F. Zhang, Face recognition from a single image per person: A survey, *Pattern Recogn.* 39 (9) (2006) 1725 – 1745. doi:10.1016/j.patcog.2006.03.013.
- [6] E. Candès, Mathematics of sparsity (and a few other things), in: *Proceedings of the International Congress of Mathematicians*, Seoul, South Korea, 2014.
- [7] D. L. Donoho, For most large underdetermined systems of linear equations the minimal l_1 -norm solution is also the sparsest solution, *Comm. Pure Appl. Math.* 59 (6) (2006) 797–829. doi:10.1002/cpa.20132.
- [8] E. Elhamifar, R. Vidal, Sparse subspace clustering, in: *2009 IEEE Conference on Computer Vision and Pattern Recognition*, 2009, pp. 2790–2797. doi:10.1109/CVPR.2009.5206547.
- [9] L. Qiao, S. Chen, X. Tan, Sparsity preserving projections with applications to face recognition, *Pattern Recogn.* 43 (1) (2010) 331–341. doi:10.1016/j.patcog.2009.05.005.
- [10] J. Yang, J. Wang, T. Huang, Learning the sparse representation for classification, in: *2011 IEEE International Conference on Multimedia and Expo (ICME)*, 2011, pp. 1–6. doi:10.1109/ICME.2011.6012083.
- [11] S. T. Roweis, L. K. Saul, Nonlinear dimensionality reduction by locally linear embedding, *Science* 290 (2000) 2323–2326. doi:10.1126/science.290.5500.2323.
- [12] X. He, S. Yan, Y. Hu, P. Niyogi, H.-J. Zhang, Face recognition using Laplacianfaces, *IEEE Trans. Pattern Anal. Mach. Intell.* 27 (3) (2005) 328–340. doi:10.1109/TPAMI.2005.55.
- [13] A. Martinez, R. Benavente, The AR face database, Tech. Rep. 24, Computer Vision Center (June 1998). URL <http://www.cat.uab.cat/Public/Publications/1998/MaB1998>
- [14] P. Y. Simard, Y. A. LeCun, J. S. Denker, B. Victorri, *Neural Networks: Tricks of the Trade: Second Edition*, Springer Berlin Heidelberg, Berlin, Heidelberg, 2012, Ch. Transformation Invariance in Pattern Recognition – Tangent Distance and Tangent Propagation, pp. 235–269. doi:10.1007/978-3-642-35289-8_17.
- [15] J.-M. Chang, M. Kirby, Face recognition under varying viewing conditions with subspace distance, in: *International Conference on Artificial Intelligence and Pattern Recognition (AIPR-09)*, 2009, pp. 16–23. doi:10.1109/ICCV.2005.167.

- [16] J. Yang, K. Zhu, N. Zhong, Local tangent distances for classification problems, in: 2012 IEEE/WIC/ACM International Conferences on Web Intelligence and Intelligent Agent Technology (WI-IAT), Vol. 1, 2012, pp. 396–401. doi:10.1109/WI-IAT.2012.46.
- [17] J. Ho, Y. Xie, B. C. Vemuri, On a nonlinear generalization of sparse coding and dictionary learning., in: ICML (3), Vol. 28 of JMLR Proceedings, JMLR.org, 2013, pp. 1480–1488.
URL <http://dblp.uni-trier.de/db/conf/icml/icml2013.html#HoXV13>
- [18] J. Yin, Z. Liu, Z. Jin, W. Yang, Kernel sparse representation based classification, Neurocomputing 77 (1) (2012) 120 – 128. doi:<http://dx.doi.org/10.1016/j.neucom.2011.08.018>.
- [19] D. Wang, H. Lu, M.-H. Yang, Kernel collaborative face recognition, Pattern Recogn. 48 (10) (2015) 3025–3037. doi:10.1016/j.patcog.2015.01.012.
- [20] J. Waqas, Z. Yi, L. Zhang, Collaborative neighbor representation based classification using l_2 -minimization approach, Pattern Recogn. Lett. 34 (2) (2013) 201 – 208. doi:10.1016/j.patrec.2012.09.024.
- [21] C.-P. Wei, Y.-W. Chao, Y.-R. Yeh, Y.-C. F. Wang, Locality-sensitive dictionary learning for sparse representation based classification, Pattern Recogn. 46 (5) (2013) 1277–1287. doi:10.1016/j.patcog.2012.11.014.
- [22] Y. Xu, X. Li, J. Yang, D. Zhang, Integrate the original face image and its mirror image for face recognition, Neurocomputing 131 (2014) 191–199. doi:10.1016/j.neucom.2013.10.025.
- [23] H. Zhang, F. Wang, Y. Chen, W. Zhang, K. Wang, J. Liu, Sample pair based sparse representation classification for face recognition, Expert Systems with Applications 45 (2016) 352 – 358. doi:10.1016/j.eswa.2015.09.058.
- [24] X. T. Yuan, X. Liu, S. Yan, Visual classification with multitask joint sparse representation, IEEE Trans. on Image Process. 21 (10) (2012) 4349–4360. doi:10.1109/TIP.2012.2205006.
- [25] Y. Yuan, J. Lin, Q. Wang, Hyperspectral image classification via multitask joint sparse representation and stepwise MRF optimization, IEEE Trans. Cybern. PP (99) (2016) 1–12. doi:10.1109/TCYB.2015.2484324.
- [26] Z. Li, J. Liu, J. Tang, H. Lu, Robust structured subspace learning for data representation, IEEE Trans. Pattern Anal. Mach. Intell. 37 (10) (2015) 2085–2098. doi:10.1109/TPAMI.2015.2400461.
- [27] Z. Li, J. Liu, Y. Yang, X. Zhou, H. Lu, Clustering-guided sparse structural learning for unsupervised feature selection, IEEE Trans. Knowl. Data Eng. 26 (9) (2014) 2138–2150. doi:10.1109/TKDE.2013.65.

- [28] A. Singer, H.-T. Wu, Vector diffusion maps and the connection Laplacian, *Comm. Pure Appl. Math.* 65 (8) (2012) 1067–1144. doi:10.1002/cpa.21395.
- [29] A. V. Little, M. Maggioni, L. Rosasco, Multiscale geometric methods for data sets I: Multiscale SVD, noise and curvature, *Appl. Comput. Harmon. Anal* (2016) in press doi:10.1016/j.acha.2015.09.009.
- [30] C. Ceruti, S. Bassi, A. Rozza, G. Lombardi, E. Casiraghi, P. Campadelli, DANCo: An intrinsic dimensionality estimator exploiting angle and norm concentration, *Pattern Recogn.* 47 (8) (2014) 2569 – 2581. doi:10.1016/j.patcog.2014.02.013.
- [31] D. L. Donoho, Y. Tsaig, Fast solution of l_1 -norm minimization problems when the solution may be sparse, *IEEE Trans. Inform. Theory* 54 (11) (2008) 4789–4812. doi:10.1109/TIT.2008.929958.
- [32] A. Y. Yang, S. S. Sastry, A. Ganesh, Y. Ma, Fast ℓ_1 -minimization algorithms and an application in robust face recognition: A review, in: 2010 17th IEEE International Conference on Image Processing, 2010, pp. 1849–1852. doi:10.1109/ICIP.2010.5651522.
- [33] C. Merkwirth, U. Parlitz, W. Lauterborn, Fast nearest-neighbor searching for nonlinear signal processing, *Phys. Rev. E* 62 (2000) 2089–2097. doi:10.1103/PhysRevE.62.2089.
- [34] C. Merkwirth, U. Parlitz, I. Wedekind, D. Engster, W. Lauterborn, TSTOOL homepage, <http://www.physik3.gwdg.de/tstool/index.html>, 2009 (accessed 6.2.15).
- [35] J. L. Bentley, Multidimensional binary search trees used for associative searching, *Commun. ACM* 18 (9) (1975) 509–517. doi:10.1145/361002.361007.
- [36] K.-C. Lee, J. Ho, D. Kriegman, Acquiring linear subspaces for face recognition under variable lighting, *IEEE Trans. Pattern Anal. Mach. Intell.* 27 (5) (2005) 684–698. doi:10.1109/TPAMI.2005.92.
- [37] M. Asif, J. Romberg, ℓ_1 homotopy: A MATLAB toolbox for homotopy algorithms in ℓ_1 -norm minimization problems, <http://users.ece.gatech.edu/~sasif/homotopy/>, 2009–2013 (accessed 31.3.2015).
- [38] A. S. Georgiades, P. N. Belhumeur, D. J. Kriegman, From few to many: illumination cone models for face recognition under variable lighting and pose, *IEEE Trans. Pattern Anal. Mach. Intell.* 23 (6) (2001) 643–660. doi:10.1109/34.927464.
- [39] AT&T Laboratories Cambridge, The database of faces, <http://www.cl.cam.ac.uk/research/dtg/attarchive/facedatabase.html>, 1992-1994 (accessed 26.3.2016).
- [40] D. N. Kaslovsky, F. G. Meyer, Non-asymptotic analysis of tangent space perturbation, *Inf. Inference* 3 (2) (2014) 134–187. doi:10.1093/imaiai/iau004.

- [41] L. Zhang, M. Yang, X. Feng, Sparse representation or collaborative representation: Which helps face recognition?, in: Proceedings of the 2011 International Conference on Computer Vision, IEEE Computer Society, 2011, pp. 471–478. doi:10.1109/ICCV.2011.6126277.